ESTONIAN UNIVERSITY OF LIFE SCIENCES
Institute of Agricultural and Environmental Sciences

Svyatoslav Rogozin

ASSESSING OF LEAF/NEEDLE LITTER PRODUCTION AND DYNAMIC

VARISE TEKKE JA DÜNAAMIKA HINDAMINE

Master's thesis
Curriculum in Environmental management and policy

Supervisor: Professor Steffen Manfred Noe, *PhD*

Tartu 2021

www.emu.ee

Eesti Maaülikool
Estonian University of Life Sciences

Estonian University of Life Sciences Kreutzwaldi 1, Tartu 51014		**Abstract of master's thesis**	
Author: Svyatoslav Rogozin		Speciality: Environmental management and policy (80407)	
Title: Assessing of leaf/needle litter production and dynamic			
Pages: 64	Figures: 18	Tables: 4	Appendixes: 2

Department: Chair of Environmental Protection and Landscape Management

Field of research and (CERC S) code: Sylviculture forestry, forestry technology (B430), Plant Ecology (B270), Statistics, operation research, programming, actuarial mathematics (P160)

Supervisor: Steffen Manfred Noe

Place and date: Tartu 2021

Plants and trees litterfall represent a primary organic carbon source in forest soils. The global forest litterfall prediction is an important research field for forestry scientists worldwide. The main research purpose of this master's thesis is the statistical estimation and prediction of past and future litterfall dynamics, based on a dataset obtained from a four-year investigation at four forest areas with different trees composition and age. To reach the main research purpose of this master's thesis, a four-year investigation litterfall dataset was analysed. The linear regression model and the generalized additive model were used for statistical analysis of the litterfall dynamics. The results demonstrates that the linear regression model is not an appropriate method to estimate litterfall detailly, but it could be used to predict litterfall trend in general. In contrast, general additive model describes and predicts litterfall process very well. Four different sample areas with different ages and trees composition were statistically estimated due to generalized additive models. All models show the differences between the forest stands, involved in this investigation. During analysis appears, that absent observations are negatively influence on both types of models. As a result, the bootstrap method was used to find the absent observations. The sampling problem

was successfully solved due it. Bootstrapped method gave an opportunity, to create new uniformly distributed observations and generate more accurate and reliable model. This study could facilitate to better understanding of the litterfall dynamics in Estonian forests.

Keywords: litterfall, linear regression model, generalized additive model, bootstrap method

www.emu.ee

Eesti Maaülikool

Estonian University of Life Sciences

Eesti Maaülikool Kreutzwaldi 1, Tartu 51014		**Magistritöö lühikokkuvõte**	
Autor: Svyatoslav Rogozin		Õppekava: Keskkonnakorraldus ja -poliitika (80407)	
Pealkiri: Varise tekke ja dünaamika hindamine			
Lehekülgi: 64	Jooniseid: 18	Tabeleid: 4	Lisasid: 2
Õppetool: Keskkonnakaitse ja maastikukorralduse õppetool Uurimisvaldkond: Metsakasvatus, metsandus, metsandustehnoloogia (B430), Taimeökoloogia (B270), Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika (P160) Juhendaja: Steffen Manfred Noe Kaitsmiskoht ja -aasta: Tartu 2021			
Alustaimestiku ja puude varis on peamine süsiniku allikas metsade muldadel. Metsa varise dünaamika ennustus on oluline uurimisvaldkond metsandusteadlaste jaoks. Käesoleva magistritöö peamine eesmärk on varasema ja tulevase varise dünaamika statistiliselt hindamine ja prognoosimine andmestiku põhjal, kus kogutud nelja-aastase uurimised neljal metsaaladel erineva puude koosseisu ja vanusega. Selle magistritöö põhieesmärgi saavutamiseks oli statistiliselt analüüsitud nelja-aastane vaarise andmekogum. Varise dünaamika statistilise analüüsimiseks oli kasutatud lineaarset regressioonimudelit ja üldistatud aditiivset mudelit. Tulemused näitavad, et lineaarne regressioonimudel ei ole sobiv meetod varise dünaamika detailseks hindamiseks, kuid seda on võimalik kasutada üldise varise tendentsi prognoosimiseks. Seevastu üldine aditiivne mudel kirjeldab ja ennustab väga hästi varise tekkeprotsess. Neli metsatükid erinevate puu vanuse ja puude koostisega oli analüüsitud üldistatud aditiivse mudeli abiga. Kõik mudelid näitasid erinevusi selles uurimises osalenud puistute vahel. Analüüsi käigus tuli välja, et puuduvad varise vaatlused negatiivselt mõjutavad mõlemale muudeli tüüpidele. Bootstrap-meetod oli kasutatud puuduvate vaatluse leidmiseks ja valimiprobleem oli edukalt lahendatud. Bootstrap-meetod andis võimalus luua uusi vaatlusi ühtlasest jaotusest ning genereerida täpsemat ja usaldusväärsemat			

mudelit. Antud töö võiks hõlbustada paremat arusaamist varise dünaamikast Eesti metsades.

Märksõnad: varis, lineaarne regressioonimuudel, üldistatud aditiivne muudel, bootstrap-meetod

TABLE OF CONTENTS

INTRODUCTION 7
1. MATERIALS AND METHODS 11
1.1 Study sites and sample areas descriptions 11
1.2 Litterfall collection 14
1.3 Sorted litterfall composition 16
1.4 Statistical Analysis 19
2. RESULTS 28
2.1 Collected litterfall dynamics 28
2.2 Outlier detection and removing process 29
2.3 Linear regression models of the litterfall dynamics 32
2.3.1 Monthly linear regression model of the litterfall dynamics 32
2.3.2 Yearly linear regression model of the litterfall dynamics 33
2.4 Generalized additive models of the litterfall dynamics 34
2.4.1 Monthly generalized additive model of the litterfall dynamics 34
2.4.2 Yearly generalized additive model of the litterfall dynamics 35
2.5 Generalized additive models for each sample area. 37
2.5.1 Monthly generalized additive models for each sample area 37
2.5.2 Yearly generalized additive models for each sample area 39
2.6 Bootstrapped generalized additive model 40
3. DISCUSSION 42
4. CONCLUSION 54
REFERENCES 56
APPENDIX 61
Appendix 1. Monthly collected litterfall box plots with detected outliers for each sample area 62
Appendix 2. Yearly collected litterfall box plots with detected outliers for each sample area 63

INTRODUCTION

A litterfall is an essential process in forest nutrient cycling (Vitousek, 1984). The litterfall result is dead organic biomass, which is available for decomposition processes. The decomposition process is the long-term process when dead organic biomass breaks apart into tiny particles until the quantity of primary biomass could not be recognised. The litterfall decomposition process consists of organic mineralisation and organic matter transformation (Krishna & Mohan, 2017).

Plants and trees litter represent a primary organic carbon source in forest soils (Novozhilov et al., 2017). At the end of the last century, some studies estimated that litter decomposition processes contribute approximately 70% of total annual carbon flux (Raich & Schlesinger, 1992). The following research confirmed that, especially in cold biomes, the decomposition process is a main component of the global carbon budget (Aerts, 2006).

Different researches around the world found a link between soil fertility (Klemmedson, 1987), soil acidity (Berger & Glatzel, 1994), climate (Starr et al., 2005), individual tree attributes (Scherer-Lorenzen et al., 2007), and decomposition of forest litter. These factors make a forest a complex ecosystem, where litter sampling is essential to understand the nutrient cycling in the forests (Vitousek, 1982). However, the amount of litterfall strongly depends on climate conditions (Lopes et al., 2015; Martínez-Yrízar & Sarukhán, 1990). Annual variations in climate and extreme weather events such as storms can dramatically influence on the dynamics of litterfall (Lodge et al., 1991). A few studies claim that a relationship between the amount of litterfall and the forest composition exists (Lowman, 1988). In the middle of the 20th century, Bray and Gorham (1964) concluded that evergreen coniferous forests produce a larger amount of litterfall than broadleaf forests. Their studies claimed that this difference is due to coniferous forest evergreen nature. Moreover, decade after this study, Millar (1974) verified Bray and Gorham (1964) conclusion. However, at the beginning of the 21st century, researchers from the Mediterranean area, temperate and subtropical areas of China refuted the previous conclusion. In addition, Lian & Zhang (1998) and Liu (2001) studies showed that broadleaf forests could have a significantly higher litterfall dynamics (Lian & Zhang, 1998; C. J. Liu et al., 2001).

In a 2004 study, Liu (2004) showed that the total litterfall in broadleaf forests is higher than in coniferous forests in temperate, subtropical, and tropical areas. However, the situation in the boreal zone is the opposite and litterfall in coniferous forests is significantly higher.

The litterfall is a continuous and endless process, but its sampling happens at discrete time points. For the future statistical analysis of litterfall, it is vital to estimate litterfall as a continuous flux of biomass. However, the accuracy of the estimation process is highly dependent on the frequency of measurements the size of the available database and sampling representativeness (Shen et al., 2019). In studying dynamics, the main aim is to observe the whole process, but practically it is not always possible. Since it is impossible it is essential to derive a representative sample (Acharya et al., 2013). The representative sample is a subset of the whole population, which is accurate in displaying parameters of the larger groups. The predictions made on a representative sample could be projected on the whole population (Henry, 2009).

The global forest litterfall prediction was an important research field for forestry scientists around the world (Lonsdale, 1988). In different studies, scientists variously predicted quantitative parameters of litterfall. In the 20th century, different statistical approaches were already used to predict litterfall parameters (Cousens & Newbould, 1968). Furthermore, due to the widely spreading of computer science in the last two decades of the 20th century, the solving of statistical tasks became more accessible (Tiit, 2016). At the same time, regression analysis is gaining wide popularity in the science (Dhakal, 2019). Regression analysis is one of the simplest and strongest statistical techniques, which provides relationships between a dependent and one or more independent variables. Regression analysis is frequently used to predict future trends and to understand the influence of factors. Different regression applications are used in the majority of fields of science: engineering, chemical and environmental sciences, economics and social sciences (Barbur et al., 1994). Because of its application to different problems, regression analysis could be the most universally used predictable statistical technique (Khuri, 2013).

In mathematical statistics, linear regression is the most commonly used type of regressions because of its simplicity in understandings and interpreting (Yan, 2008). The simple linear regression is used for modelling linear relationships between a dependent variable y and an explanatory variable X (Xin & Su, 2010).

In theoretical statistics, the observations need to be assessed so that the model gives the "best fit" to the dataset (Xin & Su, 2010). However, the theoretical part of statistics has differences from the practical part, and it usually happens when a pattern of observation is too difficult to recognise. This leads to model under- or overfitting. Underfitting occurs when the model is too simple for the dataset and cannot describe it adequately (Van Der Aalst et al., 2010). Two main reasons why a model could be very simple is a low number of features, which were implemented in the model or model is regularised too much (Kasturi, 2019). This makes the model inflexible in learning from the dataset. On the other hand, overfitting occurs when the model describes the dataset very accurate instead of showing basic relationships (Chan et al., 2011; Piotrowski & Napiorkowski, 2013). Overall, using underfitted or overfitted models for making decisions could be catastrophic for business or science studies.

Linear models are simple for interpreting and the meaning of their parameters are easy to understand. However, sometimes, it is needed more complex phenomena than linear relationships can represent (Cheng & Traylor, 1995). When linear regressions cannot describe the dataset, there is a possibility of using non-linear regression models (Amemiya, 1983). Machine learning models, like boosted regression trees or neural networks, can provide predictions of complex relationships (Vellido et al., 2012). The main problem is that they need a lot of data to train them appropriately, and their interpretation is challenging because the features they took into account may be so numerous that it is impossible to conclusively decide how the algorithm came to its result (Egger & Carpi, 2008).

Generalized additive models (GAM) could offer a middle ground between linear and non-linear models (Hastie & Tibshirani, 1990). Generalized additive models could fit complex, non-linear relationships and make good predictions in such cases. Moreover, generalized additive model results could be easily understandable and reasons of predictions very explainable (Hastie & Tibshirani, 1990). Generalized additive models are the classical appendix of general linear models in which the coefficients can be expanded as smooth functions of covariates (Hastie & Tibshirani, 1987).

Generalized additive models could be widely applied in different environmental studies. In his book, *Generalized Additive Models: an introduction with R*, Wood (2006b) explained and proved in detail the advantages of general additive models over other types of models using real examples. Moreover, generalized additive models were already successfully

implemented in litterfall studies. For example, in his studies, Edwards (2017) successfully fits general additive models to characterise trends in litterfall and climate data in tropical rainforests. His results helped to understand the dynamics of seasonal cycles in litterfall in tropical forests and demonstrated the applicability of generalized additive models in the continuous litterfall process.

The main research purpose of this master's thesis is the statistical estimation and prediction of past and future litterfall dynamics, based on a dataset obtained from a four-year investigation at four forest areas with different trees composition and age.

To reach the main purpose of this master's work, three research hypotheses were set:

1. Generalized additive models estimate litterfall as a continuous process better than linear regression models.
2. Forest stands with different ages and trees composition, which are involved in this study, have different litterfall dynamics.
3. Non-regular litterfall observations in this investigation have a negative impact on the model's prediction accuracy.

In this master's thesis is used APA 7th reference style, which is build inside Mendeley software.

Acknowledgements:
I would like to express my deepest appreciation to Steffen Manfred Noe, who is the ideal supervisor. His helpful advices encouraged me in hard times and master guidance makes a very difficult things more easier. He opened the window in the perfect world of statistical analysis and sampling for me.

Moreover, I would like to thank all SMEAR workers for their hard work. Especially, I would like to extend my sincere thanks to Beate Regine Noe, who is manually sorted samples for this study.

Lastly, I would like to thank Diana Sokurova, who is an analyst in National Institute for Health Development for her unwavering support and practical suggestions.

1. MATERIALS AND METHODS

1.1 Study sites and sample areas descriptions

This litterfall study was conducted at the Järvselja experimental forestry and training centre, approximately 40 km south-east of Tartu city, Estonia. Near the Järvselja village, the Station for Measuring Ecosystem-Atmosphere Relations (SMEAR Estonia) was opened in 2015 (Noe et al., 2015). The primary purpose of SMEAR is to assess the relation between forest ecosystems and the atmosphere and to elucidate the processes that determine how forest management influences carbon fluxes (Ezhova et al., 2018; Krasnova et al., 2019; Kulmala et al., 2020). Sample areas used in this investigation were established in previous SMEAR studies.

Trees litter was collected from four different sample areas. For each sample area was given a unique codename: A area, SP1 area, SP2 area, SP3 area. Sample areas were not randomly chosen, but each area represents a different forest stand.

Despite the lack of an accurate site description, there are previous studies which revealed that sample area A is represented by a mixture of young and old coniferous forest stands. In sample area A the dominating species is Scots pine (*Pinus sylvestris*) that form with the second abundant species Norway spruce (*Picea abies*) the main canopy. In the second level canopy few Silver birches (*Betula pendula*) are apparent. The oldest trees are approximately 170, and the youngest is 55 years old. Trees average height is about 32-35 meters in the oldest parts and about 20-23 meters in the younger stands.

Nine litter traps: A1, A2, A3, A4, A5, A6, A7, A8, A10 were established in sample area A. One part was established in the younger stand near to the SMEAR Estonia main tower, and another part in the older stand. Litter traps A1, A2, A3, A4 are located in old forest stand and traps A5, A6, A7, A8, A10 are located in younger forest stands.

The forest composition in the younger stand is very similar to sample area SP1.

Sample area SP1 is represented by young Scots pine (*Pinus sylvestris*) and Norway spruce (*Picea abies*) trees in the canopy layer with a suppressed layer of Norway spruce (*Picea abies*) and a few Silver birch (*Betula pendula)* trees. Table 1 gives a short description of stand elements and average trees heights and diameters. This dataset was collected in 2014, so new measurements need to be conducted for future studies.

Table 1. Sample area SP1 stand elements and their average values.

Stand elements		**Elements average values**	
Forest layers	**Tree species**	**Tree height (m)**	**Tree diameter (cm)**
First	Scots pine (*Pinus sylvestris*)	20,17	17,6
First	Norway spruce (*Picea abies*)	20,06	21,1
Second	Scots pine (*Pinus sylvestris*)	10,89	8,7
Second	Silver birch (*Betula pendula*)	9,67	5,3
Second	Norway spruce (*Picea abies*)	10,27	8,6
Undergrowth	Norway spruce (*Picea abies*)	5,89	5,3
Snag	Silver birch (*Betula pendula*)	3,90	5,1
Snag	Norway spruce (*Picea abies*)	8,48	7,2
Snag	Scots pine (*Pinus sylvestris*)	12,95	11,0

In the sample area SP1 were established eight litter traps: SP1 F01, SP1 F04, SP1 F05, SP1 F06, SP1 P18, SP1 F23, SP1 P32, SP1 F47.

Sample area SP2 is heavily dominated by old Norway spruce (*Picea abies*) stands. Scots pine (*Pinus sylvestris*) is co-dominating species, but only in the first layer. In the sample area SP2 has also represented some broadleaved trees such as Common aspen (*Populus tremula*), Norway maple (*Acer platanoides*), Small-leaved linden (*Tilia cordata*). In table 2 is given a short description of stand elements. In the table below, the measurements were done in 2017.

Table 2. Sample area SP2 stand elements and their average values.

Stand elements			Elements average values	
Forest layers	**Tree species**	**Proportion**	**Tree height (m)**	**Tree diameter (cm)**
First	Norway spruce (*Picea abies*)	51.2%	29,32	45,2
First	Scots pine (*Pinus sylvestris*)	48,0%	33,91	57,8
First	Common aspen (*Populus tremula*)	0.8%	26,9	27,2
Second	Norway spruce (*Picea abies*)	87,0%	20,33	25,5
Second	Scots pine (*Pinus sylvestris*)	6,5%	30,2	54,0
Second	Other broadleaved trees	4,4%	18,02	19,1
Second	Common aspen (*Populus tremula*)	1,0%	19,35	14,8
Second	Norway maple (*Acer platanoides*)	0,9%	19,08	17,3
Second	Small-leaved linden (*Tilia cordata*)	0,1%	11,20	11,6
Undergrowth	Norway spruce (*Picea abies*)	59,6%	7,58	9,1
Undergrowth	Other broadleaved trees	35,9%	6,75	8,1
Undergrowth	Small-leaved linden (*Tilia cordata*)	4,5%	10,05	9,6

In the sample area, SP2 were established eight litter traps: SP2 F01, SP2 F02, SP2 F03, SP2 F04, SP2 F07, SP2 P08, SP2 F16, SP2 F35.

In contrast with others forest stands, sample area SP3 is represented by a very young broadleaf forest. Common aspen (*Populus tremula*) is dominating in the canopy layer. Other broadleaf species are represented by, Silver birch (*Betula pendula*), Small-leaved linden (*Tilia cordata*), Rowan (*Sorbus aucuparia*). In the table below is given measurements, which also were done in 2017.

Table 3. Sample area SP3 stand elements and their average values.

Stand elements			Elements average values	
Forest layer	Tree species	Proportion	Tree height (m)	Tree diameter (cm)
First	Common aspen (*Populus tremula*)	68,0%	12,1	7
First	Silver birch (*Betula pendula*)	25,6%	9,3	3,9
First	Small-leaved linden (*Tilia cordata*)	4,0%	6,1	2,35
First	Norway spruce (Picea abies)	1,5%	2,65	1,45
First	Rowan (*Sorbus aucuparia*)	0,9%	3,6	1,3

In the sample area SP3 has established eight litter traps: SP3 F07, SP3 P13, SP3 M14, SP3 F16, SP3 P21, SP3 F22, SP3 F35, SP3 P48.

In the four sample areas, were 33 traps installed in total. All litter traps are placed more or less randomly to cover the sample plot area. Moreover, near them are located soil emission collars and the soil humidity sensors for soil-related studies.

1.2 Litterfall collection

Trees litter was collected over four years: from the ninth of June 2017 till the tenth of December 2020, which gives a good arrange of data. Totally was taken 790 samples of trees litterfall. Sampling was done approximately one time a month from each litter trap. Trees litterfall was collected approximately from the beginning of spring till the end of summer. In the winter, litterfall collections were not done because of restricted resources to handle frozen samples or a large amount of snow in the litter bags.

The litter trap is a construction made from a wooden frame and a big bucket inside this frame. The bucket is covered with non-water-resistant fabric inside, allowing water to run through it in rainy weather and drain through holes in the bottom of the buckets. The bucket diameter is 43 centimetres, and the height depends on each litter traps. The average litter trap height is 70 centimetres, but each litter trap takes into account the relief, so it stays straight. The Figure 1 is a photo of a litter trap.

Figure 1. Litter trap photo. (Author: Steffen Manfred Noe)

However, this construction has some disadvantages. The main problem is litter collection in wintertime. Estonia is located in the hemiboreal climate zone, so typically, in January and February it is snowing (Kallis et al., 2019). In wintertime, snow could get inside the trap. If the litter trap is full of snow, potential trees litter could glide across the snow cap and fall out, spoiling the accuracy of the experiment.

SMEAR station and Estonian University of Life Science workers conducted litter trap building and litter collections from litter traps. Trees litter, which falls inside the trap, was placed into special paper bags. Convenient litter trap construction gave an excellent opportunity to collect every gram of litter. Special paper bags, where tree litter was placed, were made from the special baking paper. The main reasons for using this kind of material are heat resistance and the absence of electrical conduction. Both factors could potentially spoil future litter measurement accuracy. Special bags were marked with respective litter traps.

After the previous step, non-conductive paper bags with litter inside were delivered to the Estonian University of Life Science, where is located the laboratory of plant biochemistry. In the laboratory samples were weighted on scales with the precision of 0,1 milligrams, and results were written in an Excel file. After this, tree litter was dried in a special oven at a temperature of about 70°C and weighted one more time. This step is needed to vaporise the moisture inside the litter and get dry organic material. Dried and measured trees litter was sorted into ten different fractions: leaves, pine, larch, spruce needles, cones, twigs, bark, lichen, seeds, and other plant litter, which was impossible to recognise. All fractions were weighted separately, and the results were put into an Excel database.

1.3 Sorted litterfall composition

At the fraction sorting step, it was researched that needle litter consists of two dominant species: Scots pine needles (*Pinus sylvestris*) and Norway spruce needles (*Picea abies*). European larch needles (*Larix decidua*) are also represented in the dataset, but the amount of needles is small compared with Baltic pine and European spruce needles. The presence of larch needles in litterfall composition could be explained with nearby located Taropedaja forest, where are growing European larches (*Larix decidua*). The larch needles, which are found in litterfall traps, could be a result of windy weather. Leaf litter was represented by Silver birch (*Betula pendula*), Common aspen (*Populus tremula*), Norway maple (*Acer platanoides*) leaves. Table 4 represents dry litter composition by each area.

Table 4. Collected dry litterfall composition by each sample area.

	Area A	**Area SP1**	**Area SP2**	**Area SP3**
Leaves (g)	2,02	0,89	7,89	16,02
Pine Needles (g)	57,83	280,81	0,37	0,09
Larch Needles (g)	0,03	0,11	0	0
Spruce Needles (g)	41,91	109,63	368,43	3,97
Cones (g)	23,15	39,77	5,47	0
Twigs (g)	59,9	66,87	47,72	6,49
Bark (g)	27,26	50,04	4,19	0,86
Lichen (g)	3,37	6	5,26	0,01
Seeds (g)	13,15	60,07	33,79	12,29
Rest (g)	5,54	15,12	17,52	1,99
Number of sorted samples	**30**	**71**	**65**	**19**

Table 4 shows that different sample areas have unequal numbers of sorted samples. Samples from areas SP1 and SP2 are more frequently sorted than samples from areas A or SP3. The unequal number of sorted samples is the result of the fact that sample manual sorting is a hard and time-consuming process, which takes much of human resources.

For better visualisation of Table 4, pie charts are presented in the Figure 2, where dried and sorted fractions are represented in percentages of whole collected litterfall mass in each sample area.

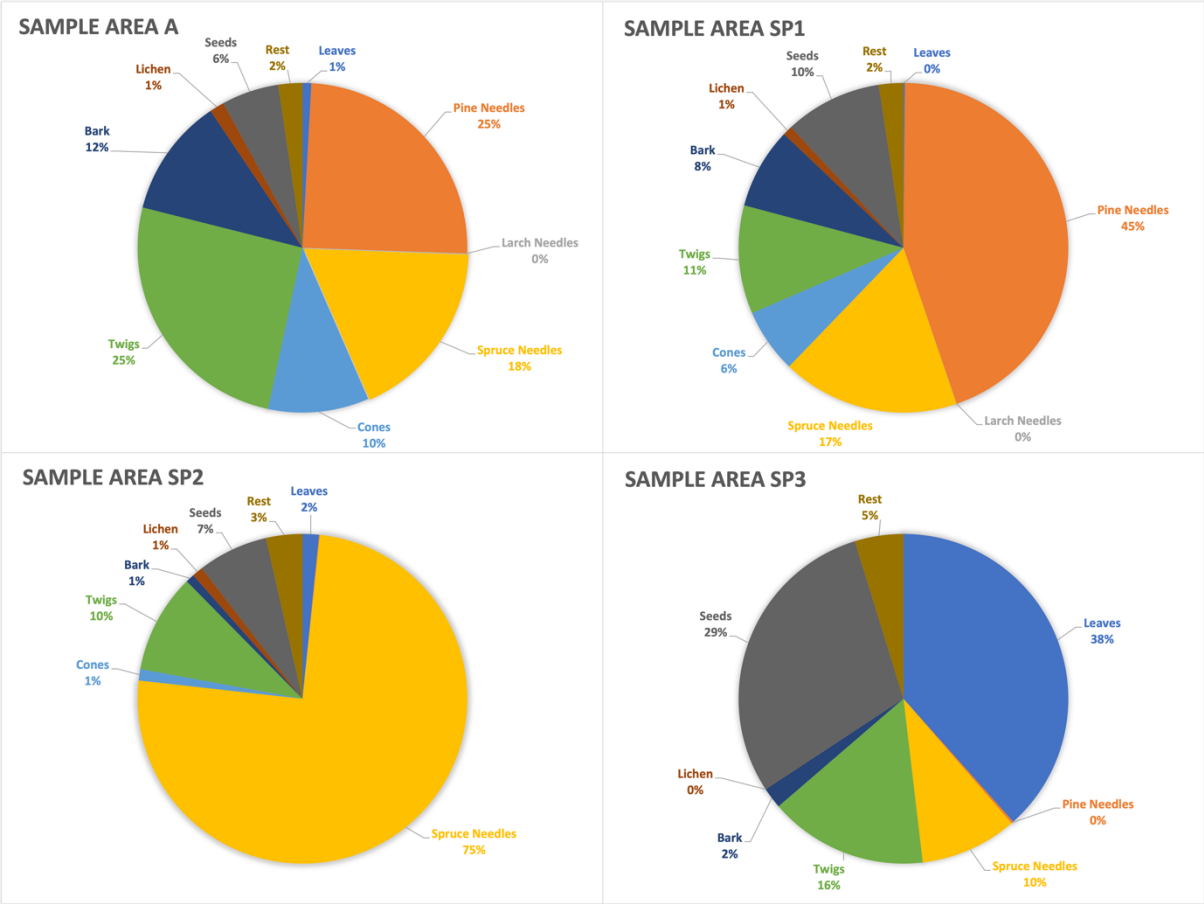


Figure 2. Sorted litterfall composition pie charts.

In sample area A, pine and spruce needles form 43% of the whole litterfall mass. Unexpectedly, the twig's part is making a quarter of the whole litterfall in this sample area. Other 32% of the whole litterfall is divided by bark, cones, seeds, and lichen. The percentage of leaves and larch needles is extremely low.

Even though sample area A and sample area SP1 have partly similar forest stand, in sample area SP1, pine needles heavily dominate over spruce needles. Together, pine and spruce needles form 62% of the whole litterfall, which is a more significant result than in sample area A. Seeds represent one-tenth of the whole litterfall, which is a higher result compared with two other coniferous forests. Other 28% of litterfall is distributed by twigs, bark pieces, cones, and rest litterfall. The number of leaves and larch needles is also extremely low.

In sample area SP2, spruce needles are heavily dominating and forms 75% of the whole litterfall. Twigs and seeds together make another meaningful part of litter composition. Leaf litter is represented by Common aspen (*Populus tremula*) and Norway maple (*Acer*

platanoides) trees. Bark pieces, lichen, larch, and pine needles are also represented, but their parts are relatively small.

In contrast with others, sample area SP3 represents broadleaf forest, where the most significant part of litterfall is leaves and seeds. Together they form 67% of the whole litterfall in this sample area. The twig's part is slightly bigger compared with sample areas SP1 and SP2. Although, a few Norway spruces (*Picea abies*) giving own impact with needles in litter composition. The amount of bark is very minor compared with coniferous forest stands. However, this sorting result could lead to an inappropriate sorting sampling. Sixteen sorted samplings were done in the period from June until July. Two more were done in April, and only one was done in November. It is expected that in June or July will be more seeds because the majority of tree species in this sample area flowering in late springtime.

Overall, sorted litterfall reflects the composition of the trees.

1.4 Statistical Analysis

To reach the main research purpose of this master's thesis, a statistical analysis was conducted using R 4.0.4 and Microsoft Excel 16.48 software.

The first statistical analysis stage is a preparation of representative sample dataset for future modelling. The raw Excel dataset was rearranged to be suitable for analysis in R software. At this step, new additional columns such as *date*, *the month of investigation*, *the month of the year*, *the day of the year*, *the day in excel format* were created. All these columns are representing different time measurement methods.

Column *date* represents the date when litterfall was collected in classical date format (from 09.06.2017 to 10.12.2020). This column is needed to conduct primary analysis and for internal calculations.

The column *month of investigation* represents the month of investigation when trees litter was collected. This column presents 43 months of study, so the first month of the study is June 2017, and the last month is December 2020. This column is needed to generate linear regression and generalized additive models.

The column *month of the year* represents the months from January till December. This column is needed to generate yearly linear regression model and generalized additive models.

The column *day of the year* represents the day of the calendar year. This column represents calendar days in scale from 1 till 365, where 1 is the first of January and 365 is thirty first of December. This column was generated with Excel YEAR function. This column is needed to generate bootstrapped daily data for generalized additive model.

The column *day in excel format* represents the day when the litter was collected as a number value. However, needed packages in R, such as the mgcv package, are not suitable to correctly analyse data with date class, so it was decided to convert the date with the Excel DATEVALUE function. Function DATEVALUE converts a date to a serial number that Excel and R could recognise. It is essential to know that dates begin from 01.01.1900, which is the first day of the Excel calendar (Microsoft Corporation, n.d.). As a result, there are days serial numbers, which can be analysed with needed packages. This column is needed for internal calculations and conversions.

After creating the new columns and filling them with data, the primary statistical analysis was conducted. This step is needed to find the occurred problems in the dataset and estimate the number of absent observations. To reach the purposes of primary analysis there were generated four-year investigation box plot in R software. The last step before modelling is removing outliers from the original raw dataset. Outliers are exceptional values in a dataset. Outliers could be a problem for many statistical analyses because they can warp real results (Frost, 2018). Outliers removing helps increase the accuracy of future models due to minimizing influence from abnormal observations. The box plot method is the easiest way to visualise, find and remove outliers (Soetewey, 2020).

The second stage of statistical analysis is generating and assessing models on the existing data. The second step main purpose is to find out verification on the first hypothesis of this master's work.

In this master's thesis were used two different types of statistical models. The first is linear regression model, which the most popular equation with m regressors to predict a regressed variable is shown below.

$$y = B_0 + B_1 X_1 + \cdots + B_m X_m + \varepsilon,$$

where

y – dependent variable,

X_i – independent variable, where $i \in \{1, \dots, m\}$,

B_0 and B_i - regression coefficients, where $i \in \{1, \dots, m\}$,

m - number of predictors,

ε - residual error.

To allow predictable modelling with linear regression, the observations must be checked for correspondence with the mathematical assumptions for linear regression. In different mathematical investigations, assumption quantity is varying. In this master's thesis will be listed assumptions based on the book *The Assumptions of the Linear Regression Model* (Poole & O'Farrell, 1971). There are four linear regression assumptions:

1. Existing linear relationships between the independent variable and the dependent variable.
2. The residuals must be independent.
3. The residuals have constant variance at every level of the independent variable
4. The residuals of the model are normally distributed.

In most cases, linear models are used when the fundamental dynamics of processes is unknown. In case of litterfall as a continuous process, the fundamental dynamics is researched. However, the quantity of litterfall is heavily influenced by spatial and temporal variability (Edwards et al., 2017).

The second is the generalized additive model, which equation is following:

$$g(\mu_i) = X_i^* \theta + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p),$$

where

$$\mu_i \equiv E(Y_i),$$

E – Mathematical expectation,

Y_i –response variable,

X_i^* – row of the model matrix for any strictly parametric model components,

θ - corresponding parameter vector,

X_i – predictor variables $i \in \{1, \dots, p\}$,

f_i – smoothing functions $i \in \{1, \dots, p\}$.

To verify or reject the first hypothesis, four models needed to be generated and statistically estimated. To describe litterfall as a continuous process that happened during four years of investigation, the variable *month of investigation* was used as independent variable in both linear regression and generalized additive models. To describe litterfall as a cyclic process, what is happening every year, the variable *month of the year* was used in both models.

Using R software were done four models:

1. The monthly linear regression model, where the independent variable is the month of investigation, and the dependent variable is collected dry litterfall.
2. The yearly linear model, where the independent variable is the month of the year, and the dependent variable is collected dry litterfall.
3. The monthly generalized additive model, where the independent variable is the month of investigation, and the dependent variable is collected dry litterfall.
4. The yearly generalized additive model, where the independent variable is the month of the year, and the dependent variable is collected dry litterfall.

For linear regression modelling were used *lm()* function build inside R software, and for generalized additive models was used *gam()* function from the mgcv package.

The monthly and yearly generalized additive models were built with the cubic splines and k-values addition.

Spline is a piecewise polynomial function. This means that the spline curve connects two or more polynomial curves. The cubic spline is a spline, which is built from piecewise third-order polynomials. The cubic spline is the best solution for the dataset involved in this investigation. Outside this master's thesis was checked the majority of popular spline types, and cubic spline demonstrates the best Akaike criteria.

K-values in the mgcv package represent the number of basis function, which is used for each smooth term before any restrictions are implemented. K-values increase the computational efficiency of penalized regression smoothers (S. Wood, 2007; S. N. Wood, 2006).

It is always an actual question how much k-values add. More k-values mean more "wiggliness" of the smoothed curve. When added a lot of k-values, the model will try to reach each data point and become overfitted. In this thesis, for the monthly generalized additive model, were added twenty k-values, and for the yearly generalized additive model, were added nine k-values. Outside this thesis was verified that this is optimal k-value quantity. Both k-value values helped improve the basic relationships between variables without heavy overfitting and increase the Akaike information criterion.

For the model's statistical estimation, in R software was used function *summary()*, which gives a short summary of generated models.

Models need to be compared pairwise using statistical parameters. The monthly linear regression model was compared with the monthly generalized additive model, and a yearly linear model was compared with the yearly generalized additive model. For choosing the best linear model were used plot method, p-value, r-squared value, Akaike information criterion.

The plot method is the easiest graphical method to estimate model fit. The main idea of this method is visually estimate model fitting to the dataset. However, when it is hard to find a visual pattern of the dataset, using the plot method could be problematic.

A p-value is used in hypothesis testing for rejecting the null hypothesis if p-value is less than 0,05 or accepting the null hypothesis if p-value is greater. In cases of both linear regressions, the null hypothesis assumes that the month of investigation or month of the year does not influence the quantity of trees litter.

R-squared is a coefficient of determination which shows a goodness-of-fit for linear models. This parameter shows the percentage of the variance in the dependent variable that the independent variables clarify mutually (Frost, 2017b). R-squared assess the relationships between the model and the predictor variable in the convenient scale of percentages (Hayes,

2020). However, a high R-squared percentage could indicate problems with models (Frost, 2017a). The main problem is that R-squared is biased by the sample sizes. If the sample size is large and exists good and a less good fitting part, the R-squared will be high, even it is visible that this does not fit (Minitab Inc, 2016).

The Akaike information criterion (AIC) is a statistical method for finding how well a model could be fitted to the data it was generated from (Bevans, 2020). The Akaike information criterion is one of the most popular and useful criteria in modelling (Cavanaugh & Neath, 2019). However, the Akaike criterion has cases when it could be useless. First of all, the Akaike criterion cannot be used to compare models with different datasets (Akaike, 1998; Brooks et al., 1989). Secondly, it is essential to use the same response variable for all models (Akaike, 1987). Moreover, if the sample size is too small, exists an opportunity that AIC will take too many mathematical parameters for a small sample and choose the best model, which is based on them. This leads to incorrect model estimation (Hurvich & Tsai, 1989).

However, for generalized additive model's statistical parameters differ from linear regression models parameters because of their non-linear features. In this master's thesis, the Akaike information criterion mainly was taken for comparing generalized additive models. For additional parameter was decided to take deviance explained parameter. Deviance is a measure of goodness-of-fit of generalized models (Lillis, 2017). In some cases, it could have a similar role with R-squared measurement or RSS score from ANOVA test (Song, 2007). However, in cases of generalized models, deviance and R-squared have different mathematical explanations (Cameron & Windmeijer, 1997), so it will be incorrectly to compare R-squared values in linear regression models, and deviance explained values.
The third stage of statistical analysis is generating and assessing generalized additive models for each sample area separately. This step is needed to validate or reject the second hypothesis of the master's thesis.

Many studies find relationships between trees age (Celentano et al., 2011; Ewel, 1976) or forest stands composition (Lebrija-Trejos et al., 2011; Norden et al., 2015) and litterfall in different climate zones. In the scope of this master's thesis was decided to generate monthly and yearly generalized additive models for each sample area, which are involved in this investigation. The possible differences between models will verify or reject the second

hypothesis of this thesis. Moreover, these models could give an additional understanding of the litterfall process in Estonian forests and gain useful results for future investigations.

From the original dataset were done four different subsets, where each represent a certain sample area. As it has already done at the previous step, outliers were removed from the dataset using the box plot method. Monthly generalized additive models and yearly generalized additive models were created in R software. Both models were generated with the mgcv package. The independent variable is the month of investigation in the monthly generalized additive model, and the dependent variable is collected dry litter. In the yearly generalized additive model, the independent variable is the month of the year, and the dependent variable is collected dry litter. Outside of this master's thesis, also, was made linear regression models for both datasets, but it became evident that the result of the linear regression model is not suitable. As a result, monthly linear regression models and yearly linear regression models were not included in the results chapter of the master's work. However, monthly linear models and their equations are used in the discussion part only to demonstrate the general trends of litterfall dynamic during the four-year investigation.

Both generated models for the sample areas compared by previously mentioned statistical parameters.

The last stage is focused on verifying the third hypothesis of this master's thesis.
A statistical estimation of the litterfall, as a continuous process with regression analysis, depends on the fulness of the dataset. The absence or non-constant frequency of observations could harm the model's prediction power. The dataset used in this investigation has missing values. It is a typical problem in statistics when some values are unknown. One of the most straightforward solutions is using the bootstrap method (Efron & Tibshirani, 1985). The bootstrap method is the resampling technique, which allows assessing statistical parameters on a whole population by sampling data with replacement (Efron, 1979). The bootstrap method allows assessing any statistical parameter using random sampling techniques (Varian, 2005). In this master's thesis, the bootstrap method is represented by the R software *runif()* function. This function generates random values of the uniform distribution (Becker et al., 1989). Uniform continuous distribution relates to incidents that are equally likely to happen and have a certain interval.

Bootstrapping applied in this thesis consist of following steps.

First, to calculate the mass in grams of collected litterfall in the day when measurements were done. Totally, were done 24 observations in 43 months of investigation.

Secondly, the average daily dry litter was calculated. For this step, was taken mass of collected litterfall from the previous step and divided on the number of days between two observations. For example, at 24.07.2017, 249,88 grams of litterfall were collected from all traps. The previous observation was done at 09.06.2017. Between these two observations is 45 days, which means that on average, in the litter trap falls ≈5,553 grams of litter daily.

However, it is incorrect to assume that daily falls constant mass of litter. The mass range of possible litterfall in this period was calculated to solve this problem. In this master's, for better calculations, the maximum mass is represented by average daily litter plus 50% more of average daily litter. For example, daily in litter trap falls ≈5,553 grams of litter. The 50% of average daily litter is ≈2,776, so the maximum mass will be ≈8,329 grams. The minimum mass is 0 grams because, in this case, mass cannot be negative.

The next important step is generating random and continuously uniformly distributed values, which quantity equals the number of days between two observations. At this step, in R software was used *runif()* function. The maximum and minimum mass of litterfall is calculated in a previous paragraph, and the number of generated observations is the number of days between two observations. For example, the number of days between observations is 45, the maximum mass is ≈8,329 grams, and the minimum is 0 grams. As a result, a function generates 45 random values using continuous uniform distribution in the interval from 0 till ≈8,329. However, this approach assumes that the sum from 45 random values will be less than really collected mass of litter. To solve this problem, how much litterfall is absent was calculated in grams and subtracted from the sum of really collected litter. After, the mass of absent litter was equally dispersed between generated values. For example, the really collected litter mass is 249,88, and the sum of generated values is 201,159. Their subtraction result is 48,721. This result is equally dividing between 45 generated elements and adding to all 45 generated values. As a result, all 45 generated values equally add 1,083 grams.

The final step is saving generated values in a new database. These steps were done 24 times because in 43 months of the investigation, were done 24 observations. However, this approach assumes that the values from the first observation are not used. For taking them into account, was decided to estimate past 31 days, what is approximately 1 month. In this case, were manually created period from 09.05.2017 till 09.06.2017, what is the month before first observation day. New values were generated and added to a new database.

As a final result, was generated random continuous uniform distributed 1321 values, where each represents a certain day in the period from 09.05.2017 till 10.12.2020 and saving the original mass of collected litter.

New generated database needed to be statistically assessed thought model generating and its describing. From previous results and calculations, it is possible to conclude that the linear regression model will not be suitable for this dataset pattern. Moreover, this assumption was checked outside the master's thesis, and conclusion was verified. In the scope of this master's thesis for this dataset will be applied generalized additive model only.

For model building and model, analysis was used column *day of the year*. Finally, the generalized additive model was generated. The independent variable is the day of the year, and the dependable variable is predictable dry litter.

2. RESULTS

2.1 Collected litterfall dynamics

For a primary analysis and a better understanding of the representative sample dataset, a box plot was generated, where the x-axis is the date, and the y-axis represents the amount of collected dry litterfall in grams. In the Figure 3 it is seen that there are many dates with absent observations, which makes the dataset non-regular. This type of dataset needs a more attentive approach in analysing than datasets with more or less regular observations.

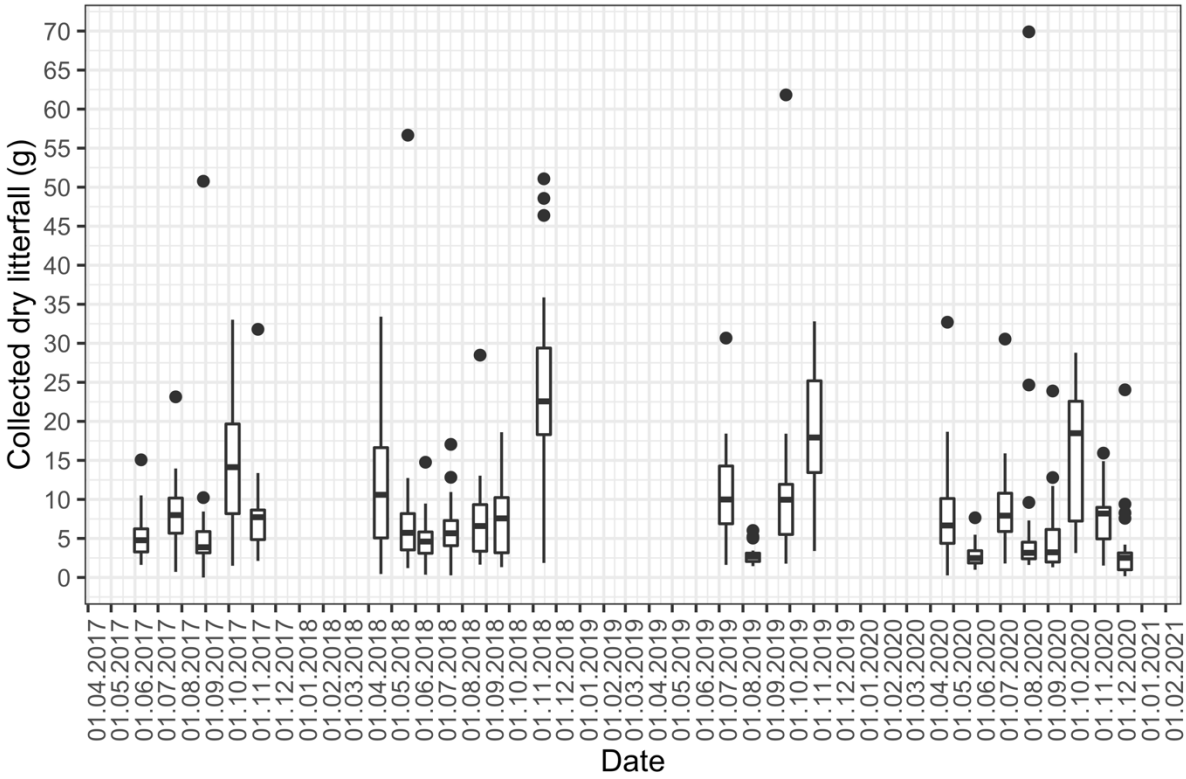


Figure 3. Collected litterfall dynamic during the whole investigation.

Despite non-regular observations, the box plot demonstrates a connection with actual natural processes. It shows that the litterfall quantity is significantly increased in the autumn, especially in October and November. In this time, in hemiboreal climate zone is observed natural litterfall, so enchanted amount of litterfall is expected. Moreover, in May and June, the litterfall amount is noticeable lower than in other months. This observation could relate to the tree's growth and reproduction time, when litterfall highly depends on weather

condition, but not on trees biology. However, strange phenomena happened in April months, when observed a large amount of litterfall compared with May months. This phenomenon will be explained in the next parts of this master's thesis.

Overall is possible to conclude that our representative sample dataset is reliable and reflect natural litterfall processes.

2.2 Outlier detection and removing process

Outliers detecting and removing process is an important task before generating models. On the one hand, removing outliers is an effective but controversial process. If humans conduct a study, exists an opportunity of making mistakes. In the litterfall study, mistakes could be made on several steps from the litterfall collection to database filling. If it is a long-term study, a single mistake's influence could not be significant, but it could seriously affect on the modelling step if it is a biased observation. Making predictions on a database with outliers could lead to incorrect results.

On the other hand, outliers removing could be unnecessary. Strange observations in environmental studies, which could be potentially defined as outliers, could be the genuine part of the process. For example, in litter trap could fall a tree branch, which mass will be significantly high. Combined with other litterfall in litter trap, mathematical algorithms could decide that observation is an outlier because the litterfall quantity of this sample is relatively high compared to other observations. As a result, removing this observation influences on reflection of the entire litterfall process. The best solution is litterfall manual sorting, which is a hard and time-consuming process. The manual sorting process gives an excellent opportunity to find out the fraction, which caused an outlier appearance.

All models presented in this master's thesis were built on datasets with outliers and datasets without them. However, outside this thesis, it was verified that outliers negatively influenced on variable relationships, so models built on datasets with outliers were not included in this work.

In this thesis, outliers detection and removing were done in ten subsets:

1. In monthly collected litterfall subset.
2. In yearly collected litterfall subset.
3. In monthly collected litterfall subset for sample area A.
4. In monthly collected litterfall subset for sample area SP1.
5. In monthly collected litterfall subset for sample area SP2.
6. In monthly collected litterfall subset for sample area SP3.
7. In yearly collected litterfall subset for sample area A.
8. In yearly collected litterfall subset for sample area SP1.
9. In yearly collected litterfall subset for the sample area SP2.
10. In yearly collected litterfall subset for the sample area SP3.

In R software, using the ggplot2 package, ten box plots were created for every ten subsets. In the Figure 4 the monthly collected litterfall box plot and the yearly collected litterfall box plot are presented. Other eight box plots with detected outliers, located in Appendix 1 and Appendix 2. Red dots on both plots are outliers. Outliers were removed from all datasets using *boxplot.stats()$out* function.

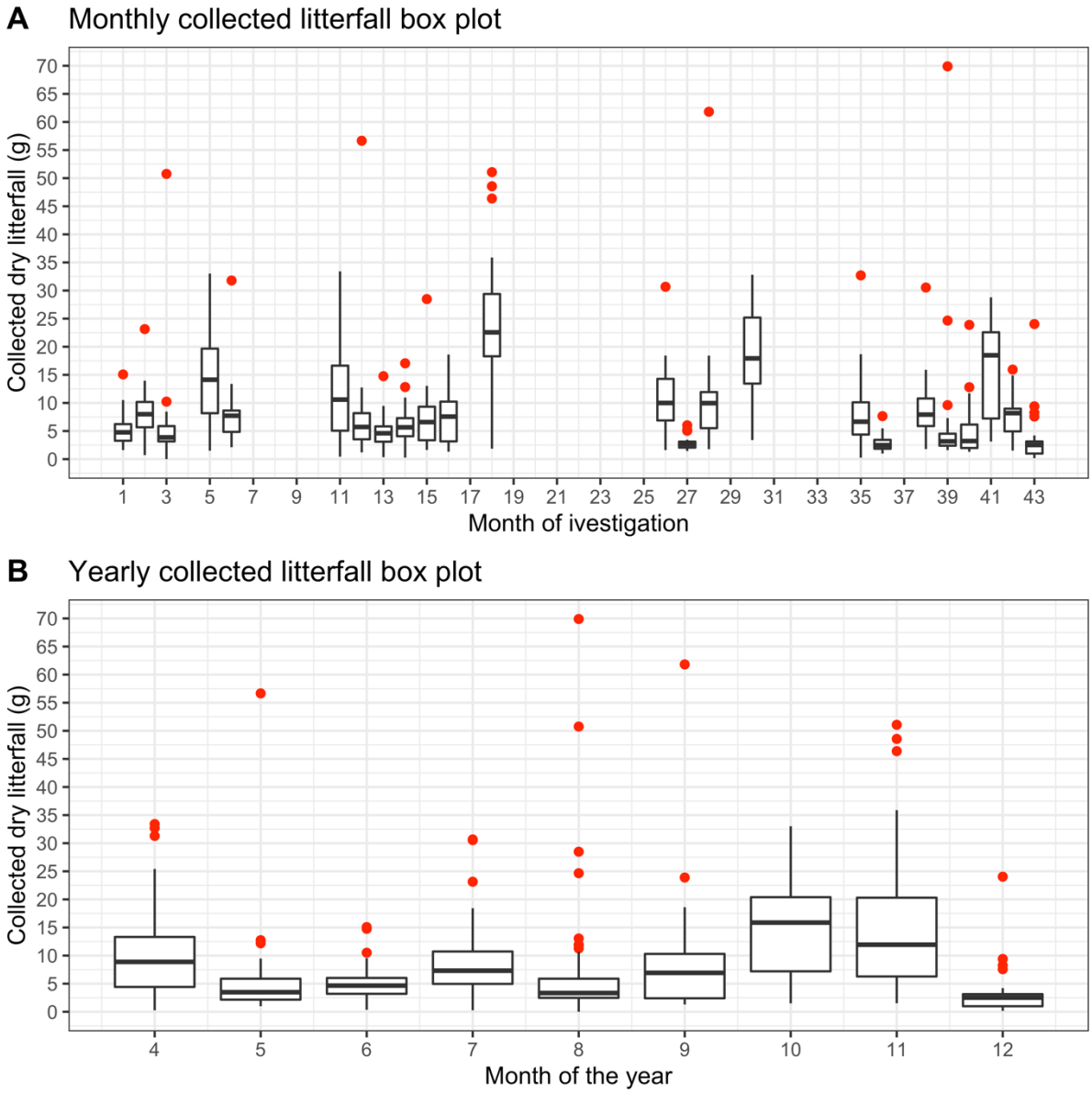


Figure 4. Monthly and yearly collected litterfall box plots with detected outliers.

Noticeable that the number of outliers on both plots is significant. From the monthly collected litterfall subset, were removed 29 samples from 790 samples, which makes 3,67% of whole observations. From the yearly collected litterfall subset were removed 27 samples, which makes 3,42% of whole observations.

2.3 Linear regression models of the litterfall dynamics

2.3.1 Monthly linear regression model of the litterfall dynamics

The Figure 5 shows a scatterplot, where the dependent variable on the y-axis is the collected dry litter in grams, and the independent variable on the x-axis is the month of the investigation. The first month of the experiment is July 2017, which is shown on the x-axis as number 1, and the last month is December 2020, which is shown on the x-axis as number 43. The dark red solid line is a linear regression model. The dark blue shaded area around the linear regression model are 95% confidence intervals.

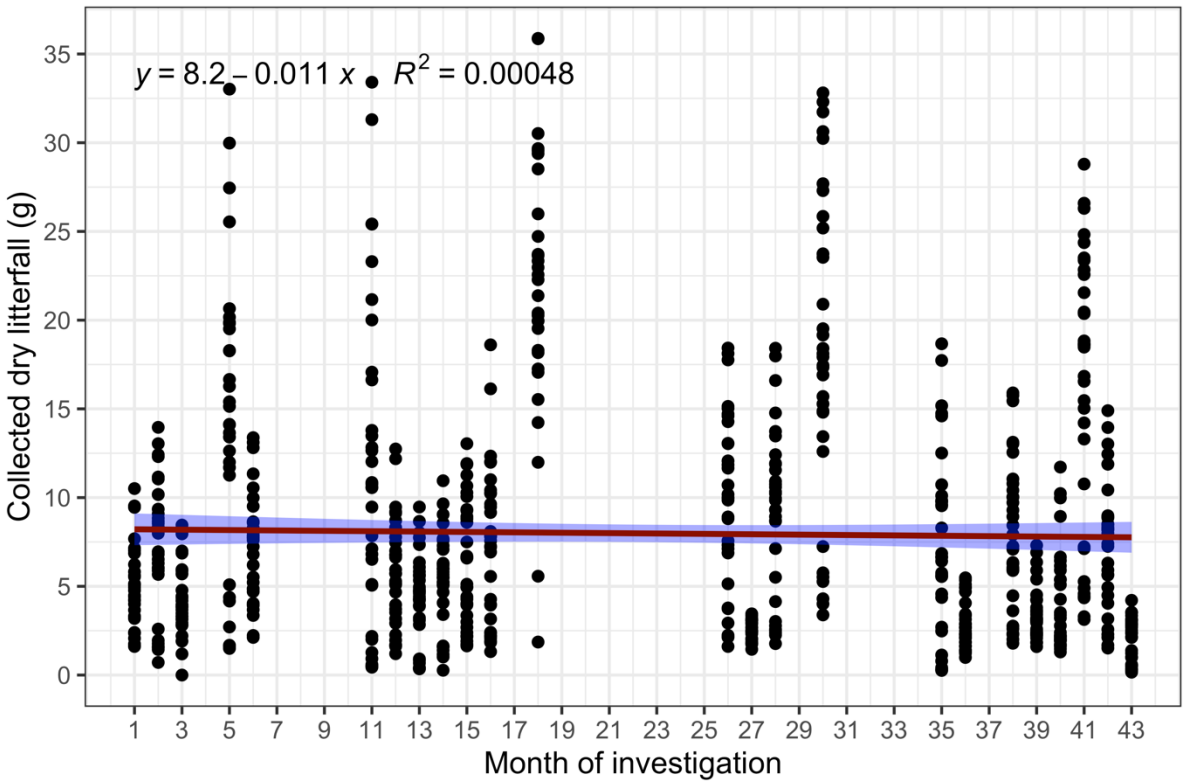


Figure 5. Monthly linear regression model of the litterfall dynamics.

From the linear regression model summary gained p-value is greater than the commonly used alpha level of 0,05. In other words, this indicates that the changes in the predictor variable are not connected with changes in the response variable. This absence of connection indicates that the month of investigation is not statistically significant, and this model cannot find the correct pattern of the dataset as it is seen in the Figure 5. Despite this result, the

linear regression equation, which is seen in the Figure 5, shows that during four years, litter quantity slightly decreased.

2.3.2 Yearly linear regression model of the litterfall dynamics

In the Figure 6 is represented the yearly linear regression model of the litterfall dynamics. The plot elements are detailly described in subchapter 2.3.1, but in this model, an independent variable on the x-axis is the month of the year from April to December.

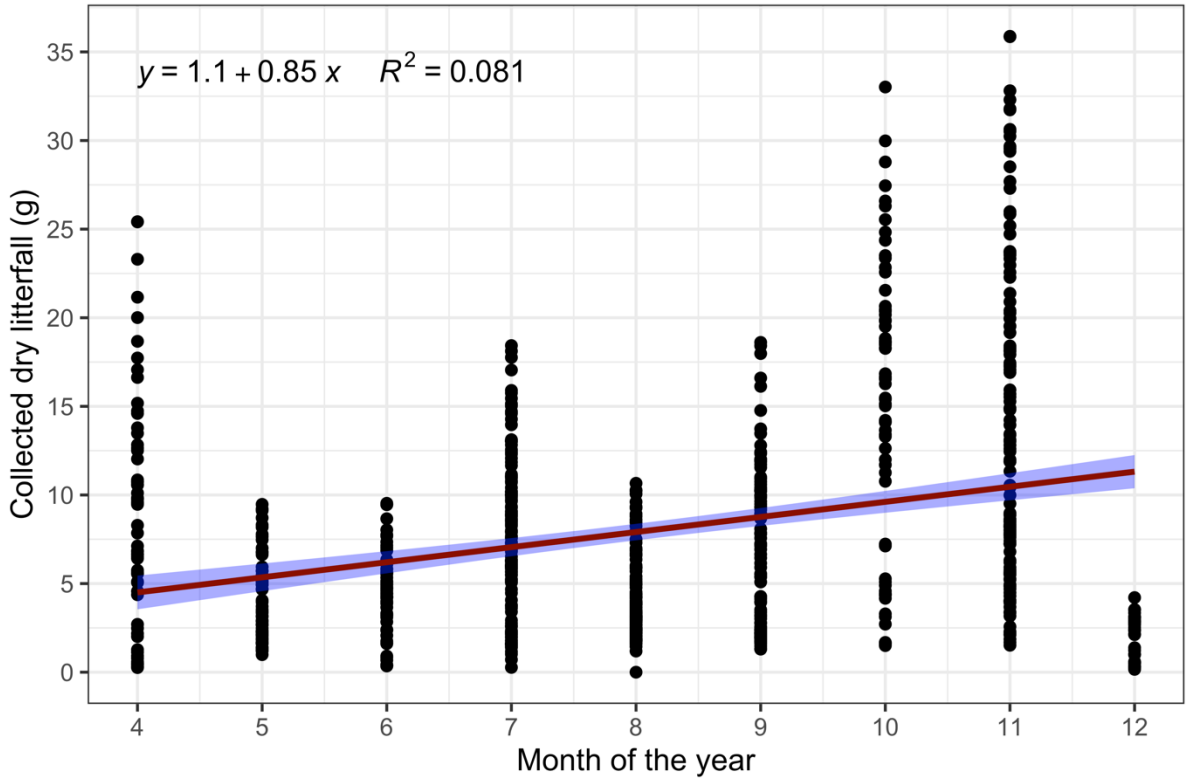


Figure 6. Yearly linear regression model of the litterfall dynamics.

The linear regression model p-value is less than the commonly used alpha level of 0,05, which means that the influence of the month of the year on litterfall dynamics is statistically significant. Linear regression equation demonstrates that litterfall quantity is constantly growing during the calendar year by 0,85 grams every next month. However, the regression model visually does not represent the real data pattern. This observation is especially visible in April and December. Implicitly, R-squared verified this fact and showed that the model covers only 8% of the whole dataset.

Finally, it is possible to conclude that this model could exist from the statistical point of view. However, this model does not show the real pattern of litterfall as a continuous process.

2.4 Generalized additive models of the litterfall dynamics

2.4.1 Monthly generalized additive model of the litterfall dynamics

In the Figure 7 is represented the monthly generalized additive model of the litterfall dynamic. The plot elements are detailly described in subchapter 2.3.1.

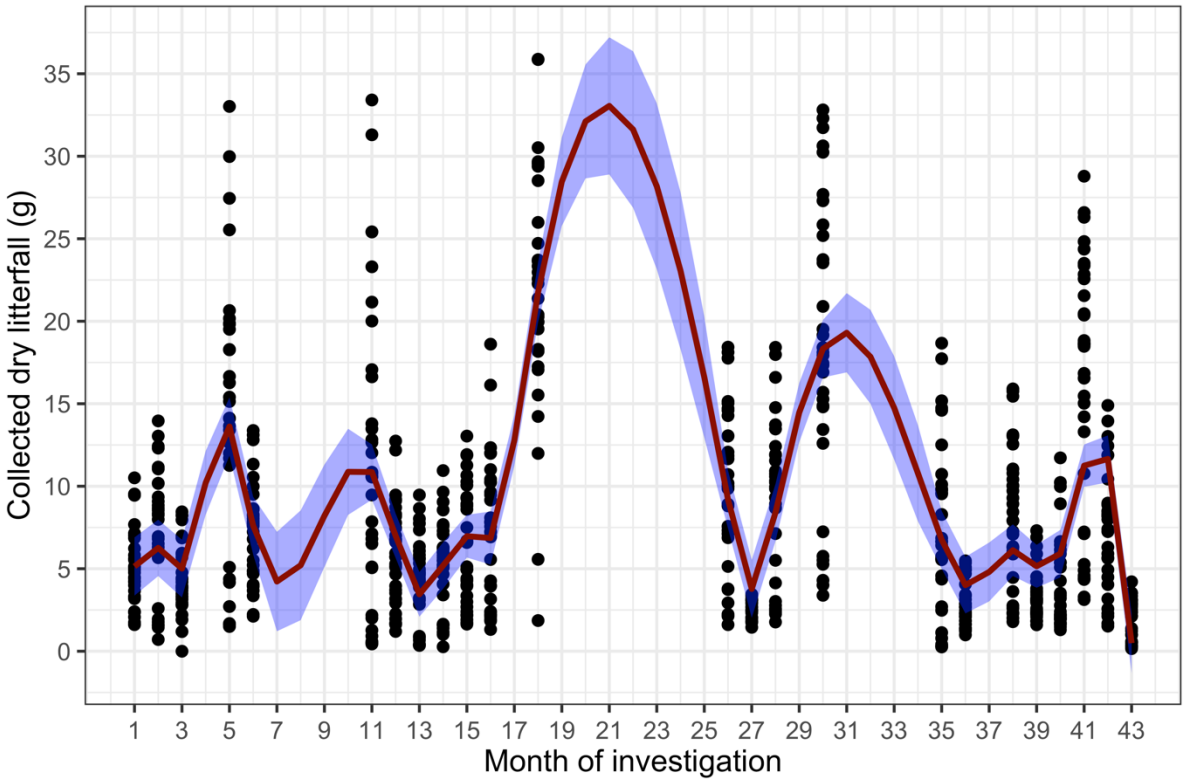


Figure 7. Monthly generalized additive model of the litterfall dynamics.

In chapter 2.3.1 was shown that the linear regression model is not suitable for this data pattern, so the generalized additive model was applied.

The generalized additive model gives more results than from the model in chapter 2.3.1. First of all, the model shows some kind of seasonality from year to year. In October, November and December litter quantity expectedly bigger. However, the lack of January, February, and March month data makes this model not as accurate as possible. For example,

in the twenty-second month of investigation, which is February of 2019, the highest spike in the whole model exists, which means that the most significant amount of litterfall was predicted in February of 2019. However, from previous studies, it is known that this spike usually happens from late September till the middle of October. This spike could be a result of model overfitting. Splines are polynomial structures, so the model fits well to the existing's points. However, model has too many degrees of freedom to relatively correct capture areas without points.

Overall, this model is demonstrating the real nature processes. However, missing data makes this model more biased in wintertime, although the deviance explained parameter equals 45,9 which is acceptable accuracy.

2.4.2 Yearly generalized additive model of the litterfall dynamics

In the Figure 8 is represented the yearly generalized additive model of the litterfall dynamics. The plot elements are detailly described in subchapter 2.3.2.

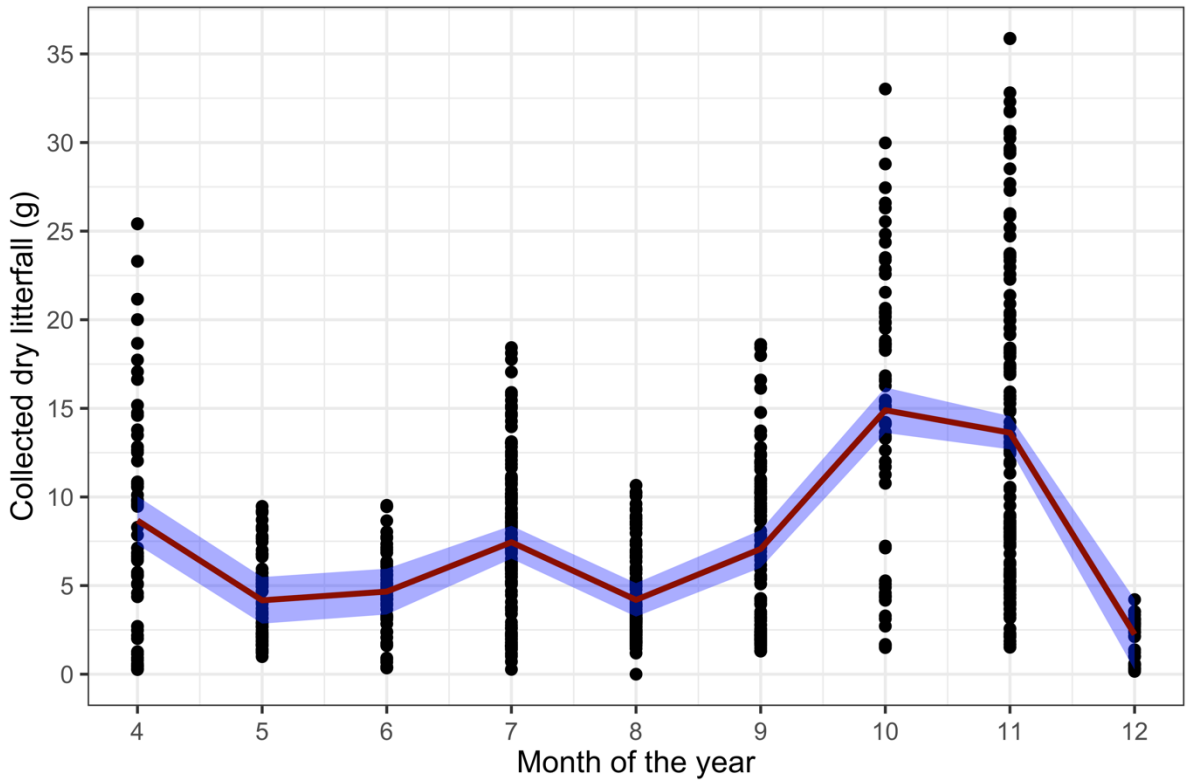


Figure 8. Yearly generalized additive model of the litterfall dynamics

In chapter 2.3.2 was proved that a linear regression model could be suitable for this dataset. However, linear regression model connection with processes in real nature is questionable. In the Figure 8 is represented the generalized additive model. Visually, this generalized additive model is representing the natural litterfall process as expected. However, the lack of wintertime data makes this model incomplete.

In April, the quantity of collected dry litterfall is significantly larger than in May. This could be a result from not being able to sample the litter traps in January, February, and March. Therefore, litter accumulated in litter traps during these months, so expectedly, the mass of litterfall in April will be relatively bigger.

After this, litterfall quantity is slightly growing from May till August. During this period happens the tree's growth process, when a lot of new biomass forms. However, new biomass is not severely damaged to be stripped off.

Next, from August till early September, litterfall is visibly growing. In this period, trees involved in this investigation are preparing for future cold weather and beginning drop litterfall on the ground.

Natural litterfall happens from the middle of September till November. As a result, litter quantity in this time is the largest. From November till December the amount of litterfall is quickly decreasing. In this period, in the northern part of the hemiboreal zone, the temperature usually decreases below 0°C, so it is essential to finish with litterfall, especially for broadleaved trees. This generalized additive model covers 34,6% of the whole dataset. From the statistical perspective, this is a dignified result.

2.5 Generalized additive models for each sample area.

2.5.1 Monthly generalized additive models for each sample area

In the Figure 9 is represented the monthly generalized additive models of the litterfall dynamics for each sample area. The plot elements are detailly described in subchapter 2.3.1.

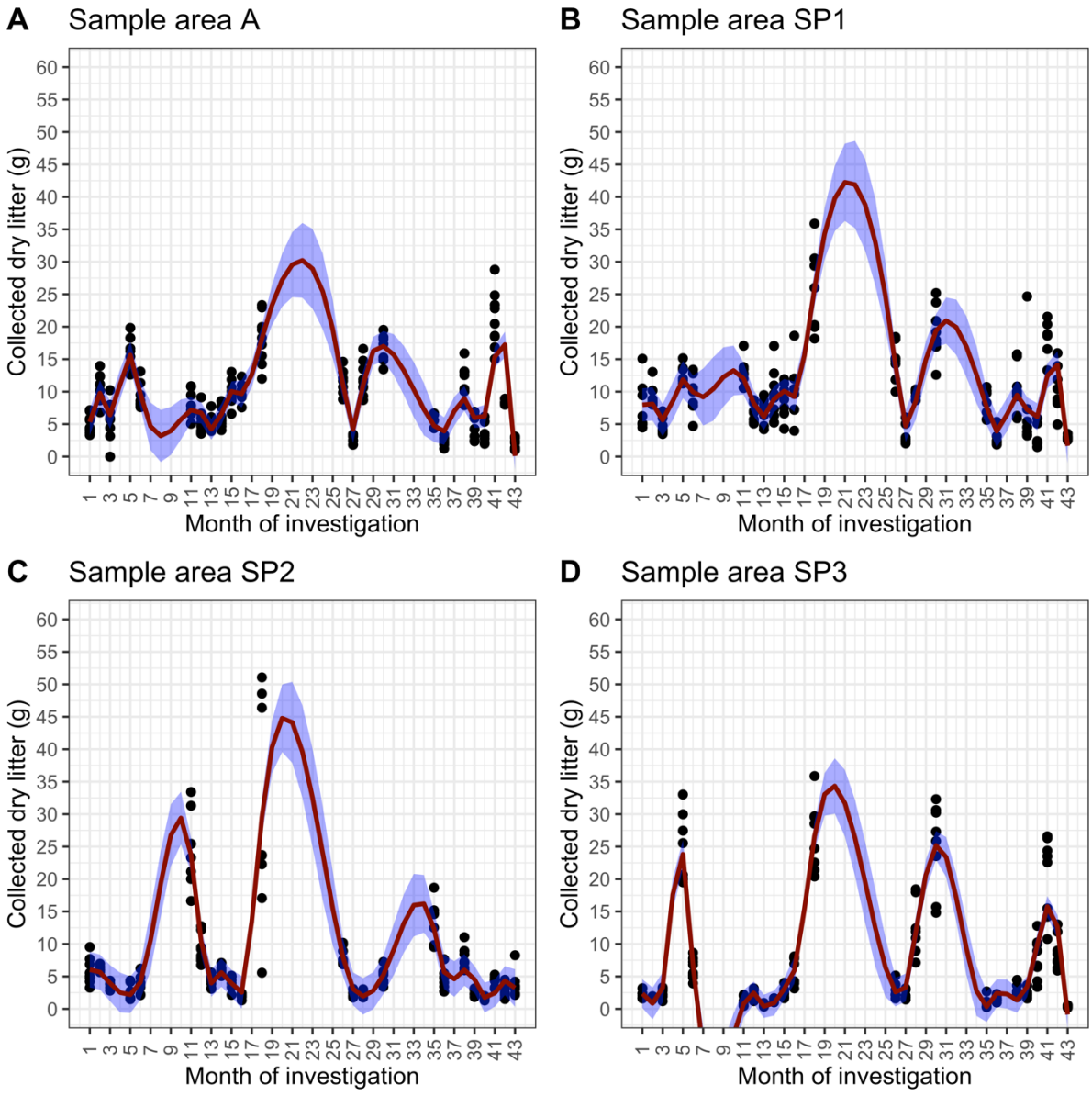


Figure 9. Monthly generalized additive models for each sample area.

Four generalized additive models were generated for better understanding the litterfall dynamics in sample areas better.

Sample area A and sample area SP1 have a very similar data pattern. The main reason is that part of sample area A (litter traps A5-A8 and A10) is the same forest stand with sample area SP1. Sample area SP1 and partly sample area A are old (about 170 years) Scots pine (*Pinus sylvestris*) stands, growing on a fen. Despite this fact, from comparing the SP1 area model and A area model is possible to conclude that litterfall overall quantity was greater in sample area SP1. Moreover, the period from the seventh (December 2017) till the tenth (March 2018) months of the investigation was predicted differently. In this period, at sample area A, the finding the exact litterfall prediction is more complicated than in sample area SP1 due to variation.

Sample area SP2 has a different data pattern compared with previous coniferous forest stands. It is noticeable that the generalized additive model has only three spikes of litterfall. In the last autumn of the investigation, litterfall quantity was significantly lower than in other years in the same autumntime.

Sample area SP3 model is visually different from others. This is a single model, which predictions fall below zero from the seventh (December 2017) till the tenth (March 2018) months of investigation. The generalized additive model also aspires to zero value in other winter months of investigation but not cross it. It is possible to interpret that the opportunity to find litterfall in the broadleaved forest is low, but it was extremely low from December 2017 till March 2018. Overall, visually this model has fewer enormous spikes and shows regular seasonality in young broadleaved forest.

Finally, different forest stands demonstrate different generalized additive models. However, in sample areas, there is the same problem when the spikes in winter months could be a result of model overfitting.

2.5.2 Yearly generalized additive models for each sample area

In the Figure 10 is represented the yearly generalized additive models of the litterfall dynamics for each sample area. The plot elements are detailly described in subchapter 2.3.1.

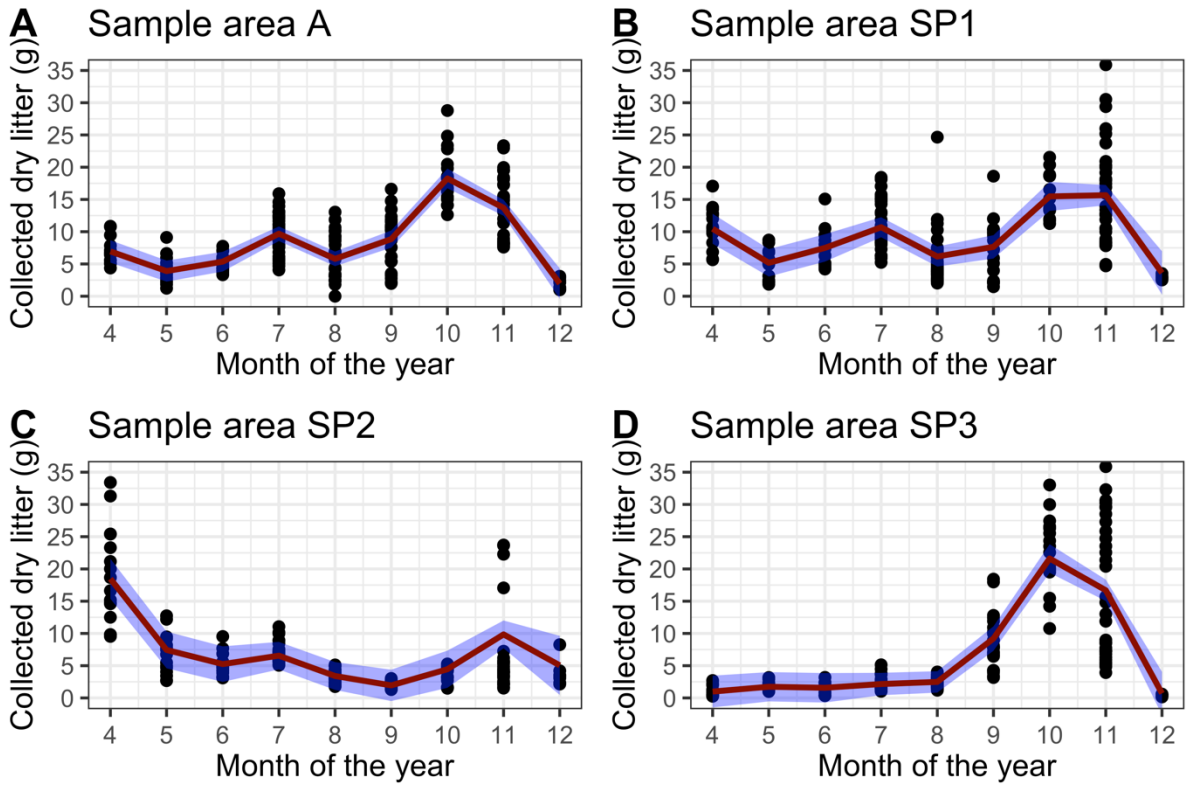


Figure 10. Yearly generalized additive models for each sample area.

From the Figure 10 follows that sample area A and sample area SP1 plots verify the results from previous subchapter 2.5.1. The overall litterfall pattern in both areas is very similar. However, one difference exists. In the Figure 10A, the litterfall peak is clearly seen in October, but in the Figure 10B, the peak is equally divided between October and November. In litterfall peak time, litterfall quantity is slightly higher in the sample area A. Lastly, the amount of litterfall is decreasing on both models in December, when the natural litterfall process is over.

Sample area SP2 differs from the other two coniferous forest stands. The most significant difference is the high amount of litterfall in April. The reasons of remarkable litterfall mass in April was described earlier, but in the sample area SP2 old and tall Norway spruces (*Picea abies*) accumulated over winter significantly more needles than in the other two coniferous forest stands. In litterfall peak time, the model demonstrates a relatively small amount of

litter compared with other coniferous forest stands. Moreover, litterfall peak time is shifted on November, which is significant dissimilarity.

Sample area SP3 has a very representative litterfall model for the broadleaved forest. First of all, constantly very low amount of litterfall is observed from April till August. After a constantly low litterfall period, begins serious and quick growth of litterfall. In October, litterfall reaches the peak, which is the most significant compared with all other three forest stands. Finally, the intensity of litterfall decreases in November, and rapid decrease of the litterfall happens in December.

2.6 Bootstrapped generalized additive model

In the Figure 11 is represented the bootstrapped generalized additive models of the litterfall. In this figure, the y-axis is bootstrapped values of the dry litter in grams, and the x-axis is the calendar day of the year. Each black dot on the Figure 11 is the predicted daily amount of litterfall. The model elements are described in subchapter 2.3.1.

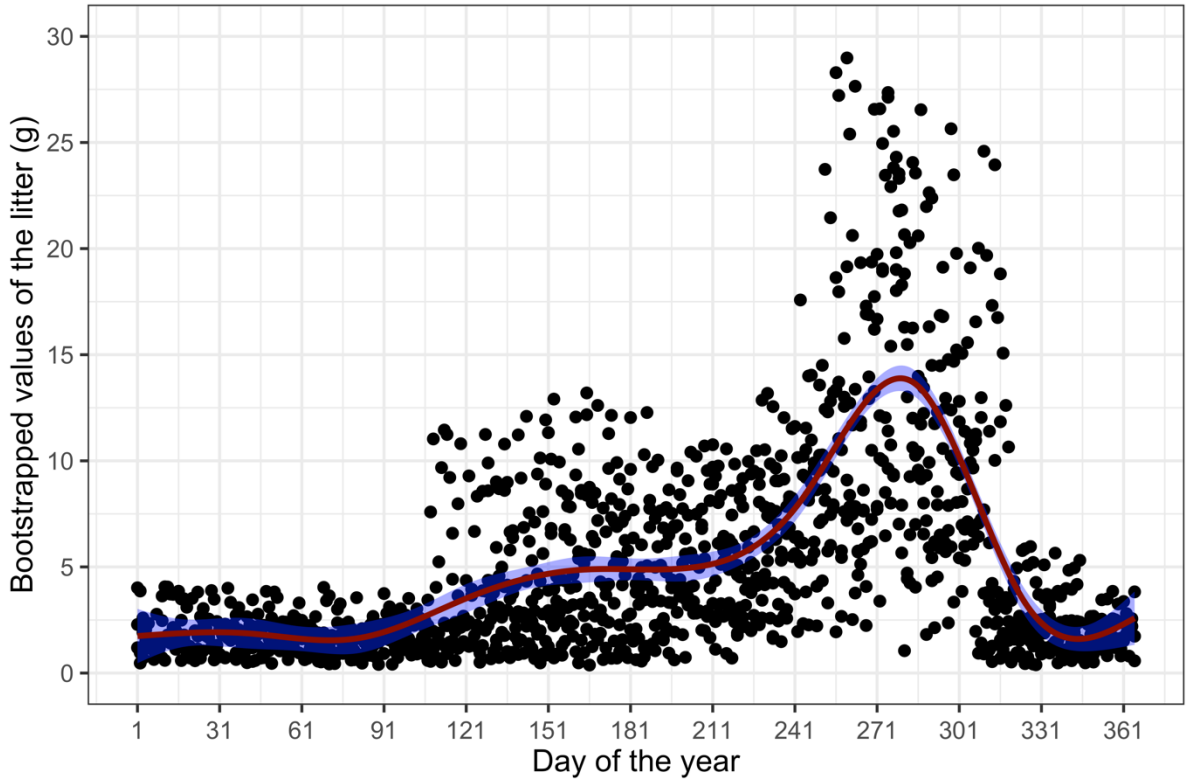


Figure 11. Bootstrapped generalized additive model.

After generating a new dataset with bootstrapped values, the generalized additive model was built. This model describes the continuous litterfall process that happened for four years. The main and essential difference between this model and the other models in this master's thesis is the presence of January, February, March data due to the application of the bootstrap algorithm.

The model demonstrates that litter amount is relatively low and more or less on a constant level from January till March. Small curves at this period could be mathematical algorithms inaccuracy, but the basic trend is shown as expected, compared with the natural litterfall process. The amount of litter noticeable increased in April, which is the result of forming new leaves and flowering processes. In May, June, July amount of litter stayed constant. However, in late August, the amount of litterfall began growing very quickly and approximately in late September, or the middle of October, peaked. This intensive growth is directly connected with the natural litterfall process. From late October and the beginning of November, the litterfall amount quickly decreases. This decrease could relate to the observation that trees, especially broadleaves trees, have already lost most of their leaves. According to this model, litterfall growth also happened in December, which could be a result of stormy weather or early snow influence. However, it is not excluded that this could be an inaccuracy of model algorithms.

K-values were not implemented in this model. It was done for two reasons. Firstly, the difference in the Akaike criterion between the model with k-values and the model without them is very low. Akaike criterion for the model without k-values is 7147,547 and with k-values is 7143,956. Secondly, deviance explained parameter without k-values, and deviance explained with maximum k-values is practically the same. Deviance explained parameter for the model without k-values is 49,8%, and with k-values is 50%.

Furthermore, in models is used cubic splines as spline type. Outside this master's thesis was checked the majority of popular spline types and cubic spline is the best variant. This result was verified by comparing different models with different spline types with the Akaike criterion.

3. DISCUSSION

To better understand the differences, the model results from subchapters 2.3.1 and 2.4.1 were compared on the Figure 12. The solid blue line represents the monthly generalized additive model, and the solid red line represents the monthly linear regression model. The dark blue shaded area around the generalized additive model and dark red area around the linear regression model are corresponding 95% confidence intervals.

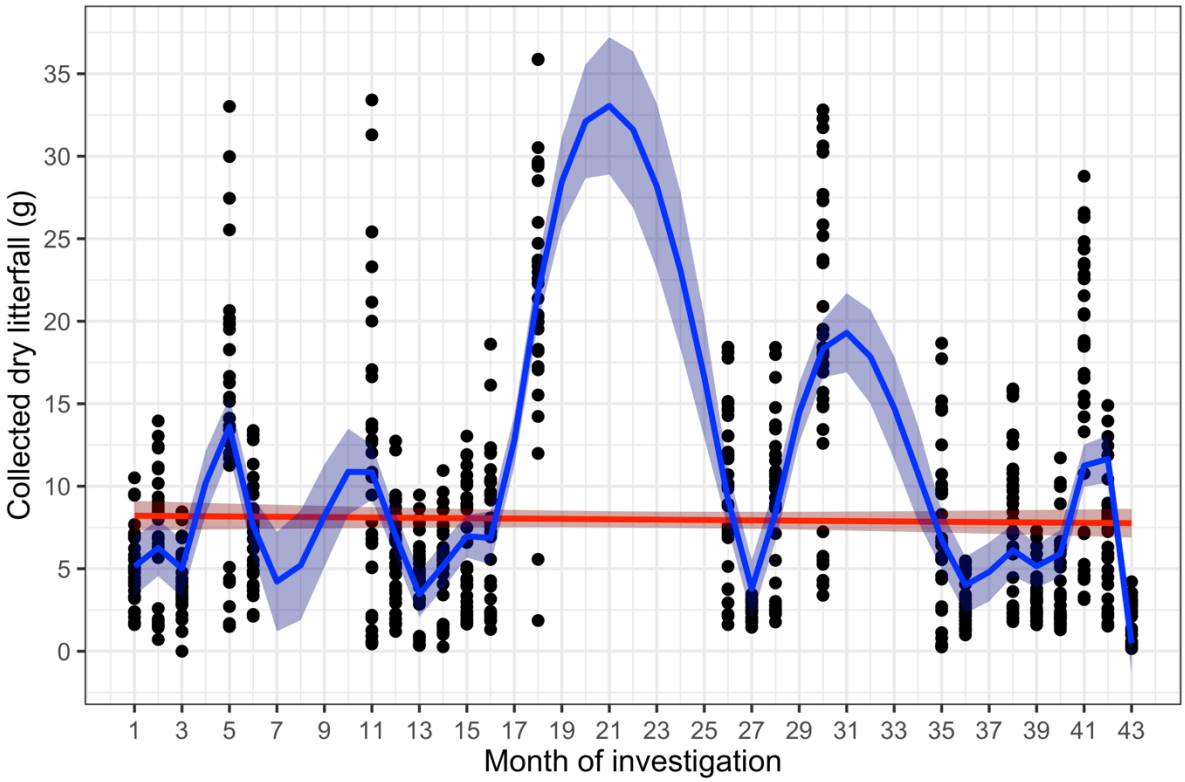


Figure 12. Compared monthly generalized additive model and linear regression models.

From subchapter 2.3.1 it became obviously, that the linear regression is not suitable to describe the dynamics in this dataset. In this type of dataset, the line and its confident intervals cannot cover much observation. Even though this model is not suitable, what is implicitly verified by this model's very low R-squared coefficient. R-squared is 0,0004828, what is coverage 0,04828% of the whole dataset. All factors together lead to underfitting using linear regression. As a result, the null hypothesis is valid, and the time parameter month of investigation does not influence the quantity of trees litter. A linear regression model could be used only as a first guess to see a possible long-term trend of the investigation.

On the other hand, the generalized additive model approach is significantly better in capturing the dynamic changes. Compared with the linear regression model, it is more flexible due to splines and k-values (20). Despite splines and k-values, the sampling problem leads to the gaps. As mentioned in subchapter 2.4.1, the computer model cannot correctly predict missing values. Generalized additive model, as all models are built on minimizing the squared distance to the dataset's mean values. If monthly observations are absent, it takes the next month with observations and is bridging the gap between these two observations. Removing outliers is an effective procedure to minimize this influence, but in this case, it helps insufficiently. As a result, the peak of litterfall moved in wintertime, which is controversial to the actual natural processes. The most prominent example is the 21st month of investigation in the Figure 12, which is February of 2019.

From a statistical viewpoint, the generalized additive model is relatively accurate and could be accepted. Akaike score of the linear regression model is 5087,962, and the generalized additive model is 4657,074, which is a significant difference. Moreover, the deviance explained parameter is 45,9, which could be interpreted as 45,9% coverage of the whole dataset.

Overall, the generalized additive model is a preferable choice in describing this dataset.

When combining all data to build a yearly dynamics (Figure 13), the yearly linear regression and the yearly generalized additive models can be compared. The plot elements are the same as in the Figure 12.

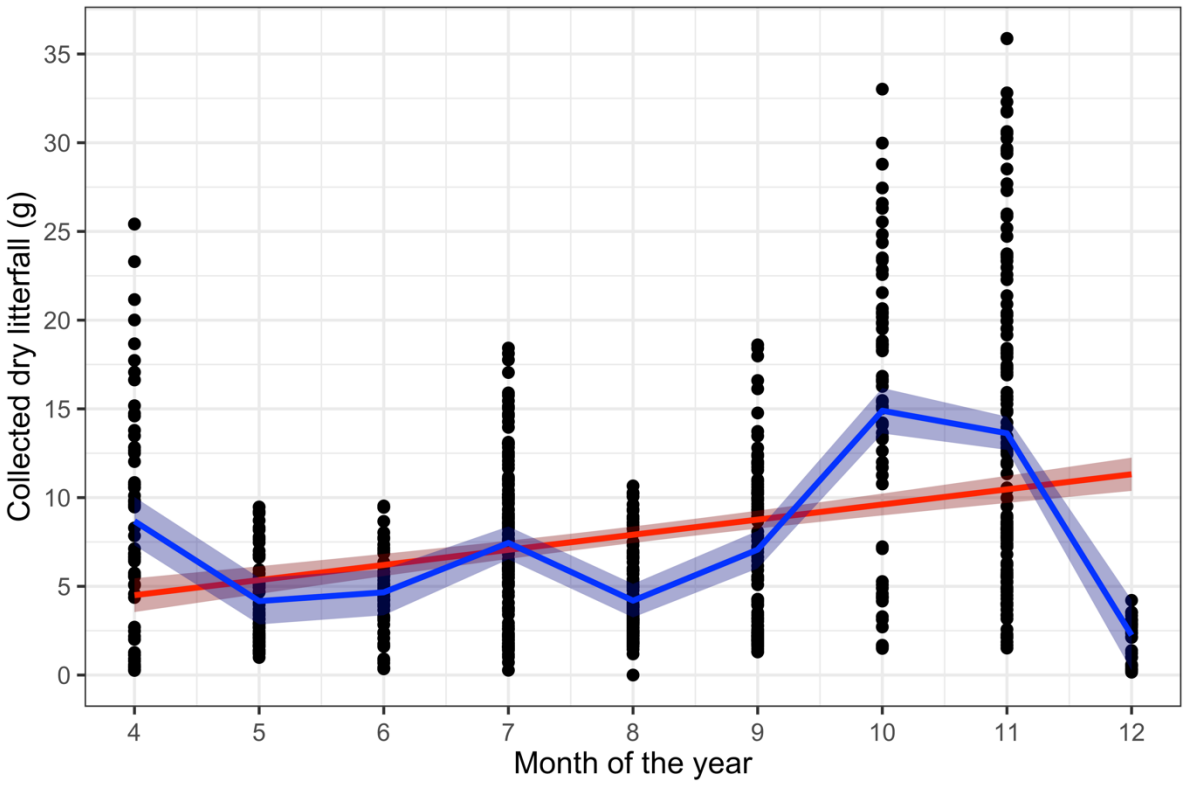


Figure 13. Compared yearly generalized additive model and linear regression models.

The yearly linear regression model's p-value is lower than the standard alpha level of 0,05. As a result, this makes the null hypothesis rejected. In addition to linear regression equation, visually is possible to notice that the model demonstrates constant litterfall growth. The possible reason is in modelling algorithms. As mentioned in the previous paragraph, all models are based on the squared distance to the mean values. At the beginning of the year the linear model is taking the low mean values in May and June and the high values in October and November into account. In this case, outliers removing was a vital process. Outside this master's thesis were compared models with and without outliers. The result was very different. The dataset with outliers demonstrates a low R-squared score of 0,01 and a steeper line slope. The dataset without outlier used in this master's thesis shows an eight-time better R-squared score, which means that model covers 8% of the whole dataset.

However, a constant growth of litterfall over the season is not an appropriate description of the actual litterfall process during the year. This model could be used in some preliminary studies, but using it in fundamental studies is not recommended.

On the other hand, the generalized additive model shows flexibility and good description possibilities. It shows a constant level of litterfall, with slight growth in July, from May till August. From August till October the amount of litterfall rapidly increases with the peak in October. In November, the amount of litterfall is slightly lower than in the peak month of October, and from November till December, litterfall quantity is abating. This model describes the actual process and verifies results from other studies. However, one problem place exists. From April till May generalized additive model shows decreasing processes, which is not typical in hemiboreal climate zone forests. The main reason for this observation is that in April, the litter amount in litter traps is larger because previous litter collection was done in November or December. From previous litter collection, litterfall is accumulating inside the trap, and, as a result, in the next collection, which happens in April, the amount of litter will be noticeably larger.

From a statistical point of view, the generalized additive model has deviance explained score of 34,6, which is not a very significant result. The reason is the time measuring approach, which brings an elongated column of data points. Compared with the yearly linear regression model with a 5024,782 AIC, the yearly generalized additive model has 4778,865 AIC, which is a 5% better result.

In conclusion, the linear regression model could be used from a statistical point of view, but the connection with the actual litterfall process is questionable. This model is suitable for showing the general trend of litterfall mass but not for describing or making predictions. On the other hand, the generalized additive model is more representative in describing and predicting the litterfall process. Due to their flexibility, they allow finding and predicting spikes and reductions in the dataset and making more or less accurate monthly litterfall predictions.

The previously used subsets consist of observations of different sample areas. However, sample areas composition is different, therefore it will be reasonable to estimate the litterfall dynamic in details for each sample area separately. Following the previous results, the linear regression model and its equation will only be used to detect general trends and general additive models for more detailed analysis.

In the Figure 14 is represented two generalized additive models for sample area A. The sample area is a mixture of old and young forest stands, where Scots pines (*Pinus sylvestris*) and Norway spruces (*Picea abies*) dominate.

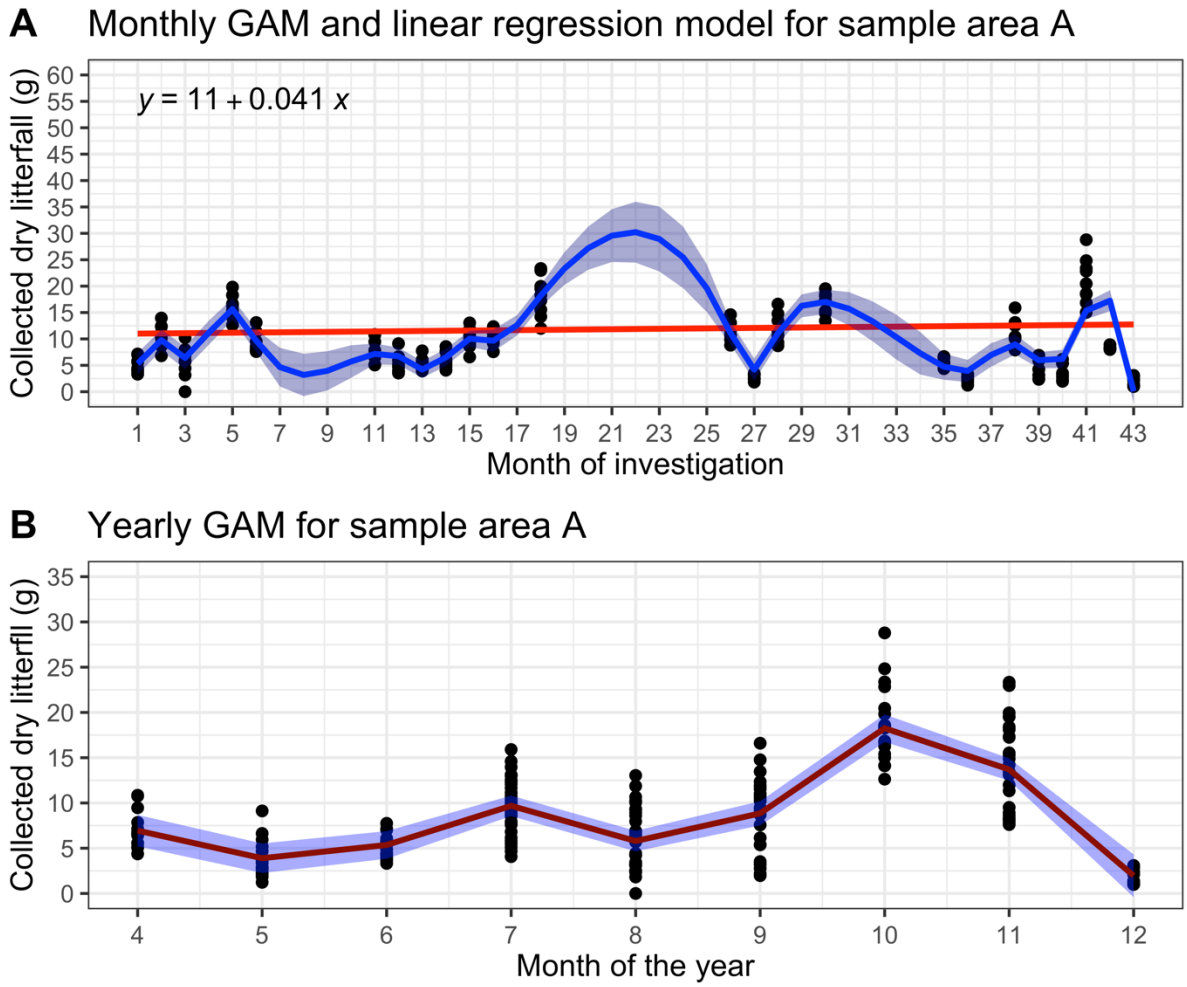


Figure 14. Monthly and yearly generalized additive models for sample area A.

The monthly generalized model not perfectly, but in general, shows the basic relationships between the collected litterfall and the months of study. The deviance explained parameter is 71,2, which is a significant result. However, this high result was achieved, by adding the k-value parameter, so the overfitting effect is presented. This overfitting effect could be noticeable in moths with absent observations, especially from the 19th till the 25th months of study. The linear regression equation, which is presented in the Figure 14A, shows that litterfall quantity is slowly growing during four years of investigation. The possible reason for this observation is the presence of most litter traps in the young forest stand while only four are in old stands where growth is slower.

The yearly generalized additive model shows the expected litterfall process during one calendar year. This model has a strongly marked period of forest growth from April until August when the amount of litterfall is not significantly varying and natural litterfall in October when the amount of litterfall is the largest during the year. The accumulation of litterfall during months with absent data is not very large, which means that mixed forest's litterfall is not significant in wintertime. The deviance explained parameter is 63,3, which is a good result. Overall, yearly GAM is not hardly overfitted or underfitted and demonstrates the basic relationships between variables.

As it was previously mentioned, sample area SP1 have similar trees composition with sample area A. In the Figure 15 is represented monthly and yearly generalized additive models for sample area SP1.

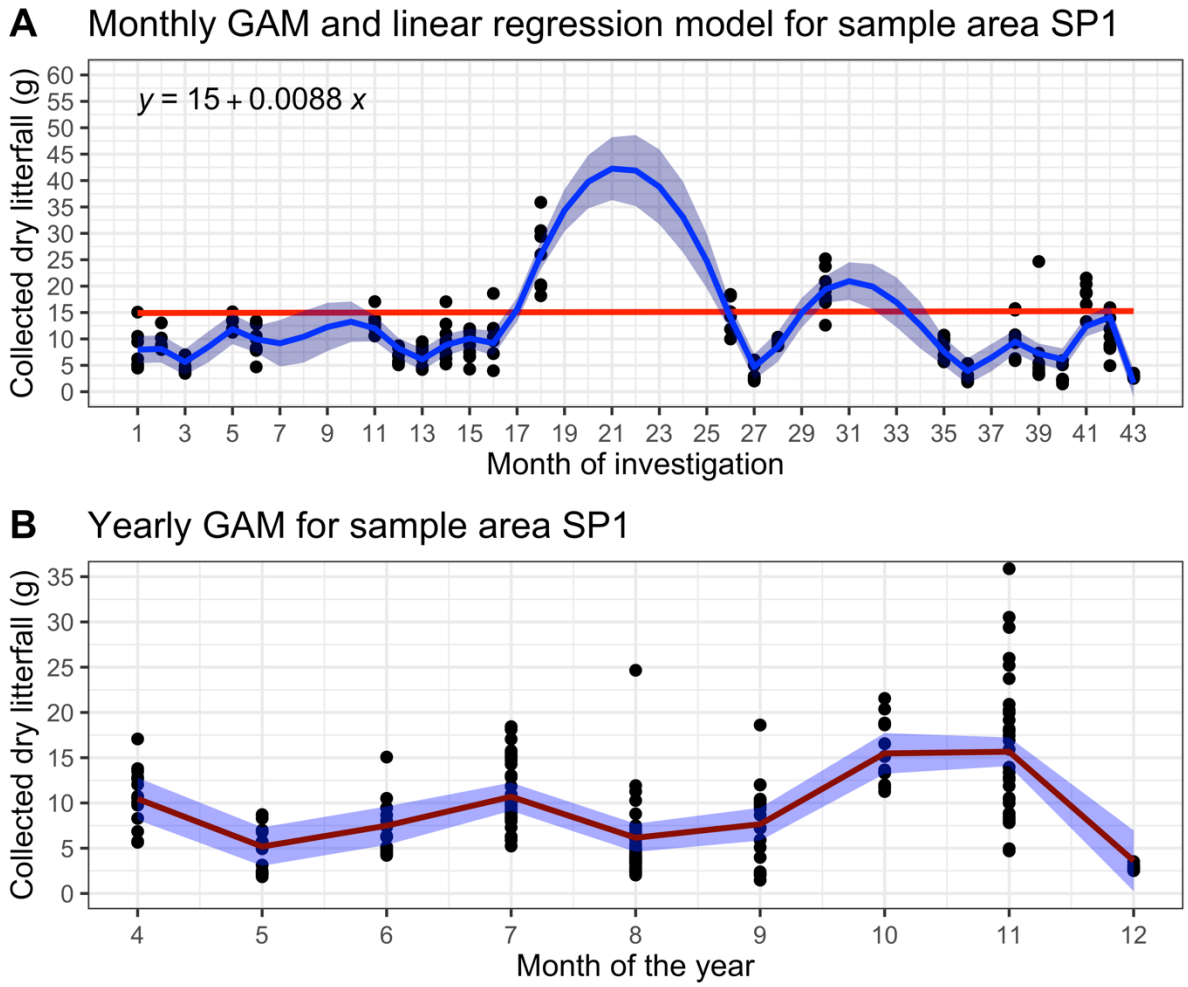


Figure 15. Monthly and yearly generalized additive models for sample area SP1.

Both generalized additive models are very similar to previously described sample area A models. However, a few differences between monthly generalized additive models exist. In sample area SP1, during four years of investigation, the litterfall amount is also slowly growing but significantly slower than in sample area A. The deviance explained parameter is also similar to sample area A and is 67,5, which is a high result.

Likewise monthly GAM, the yearly GAM is very similar to the sample area A model. However, the significant difference is the lack of a strongly marked litterfall peak. Litterfall peak is more or less equally divided between October and November. Unexpectedly, the yearly generalized additive model deviance explained parameter is only 45,4, significantly lower than area A result. This could be explained by higher variation in November and existing extreme values in August and September.

Overall, both models demonstrate good basic relationships between variables.

Sample area SP2 is also a coniferous forest, but its trees composition is very different from the previous two forest stands. In the sample area SP2, is heavily dominating old spruces in different forest layers. In the Figure 16 is represented monthly and yearly generalized additive models for sample area SP2.

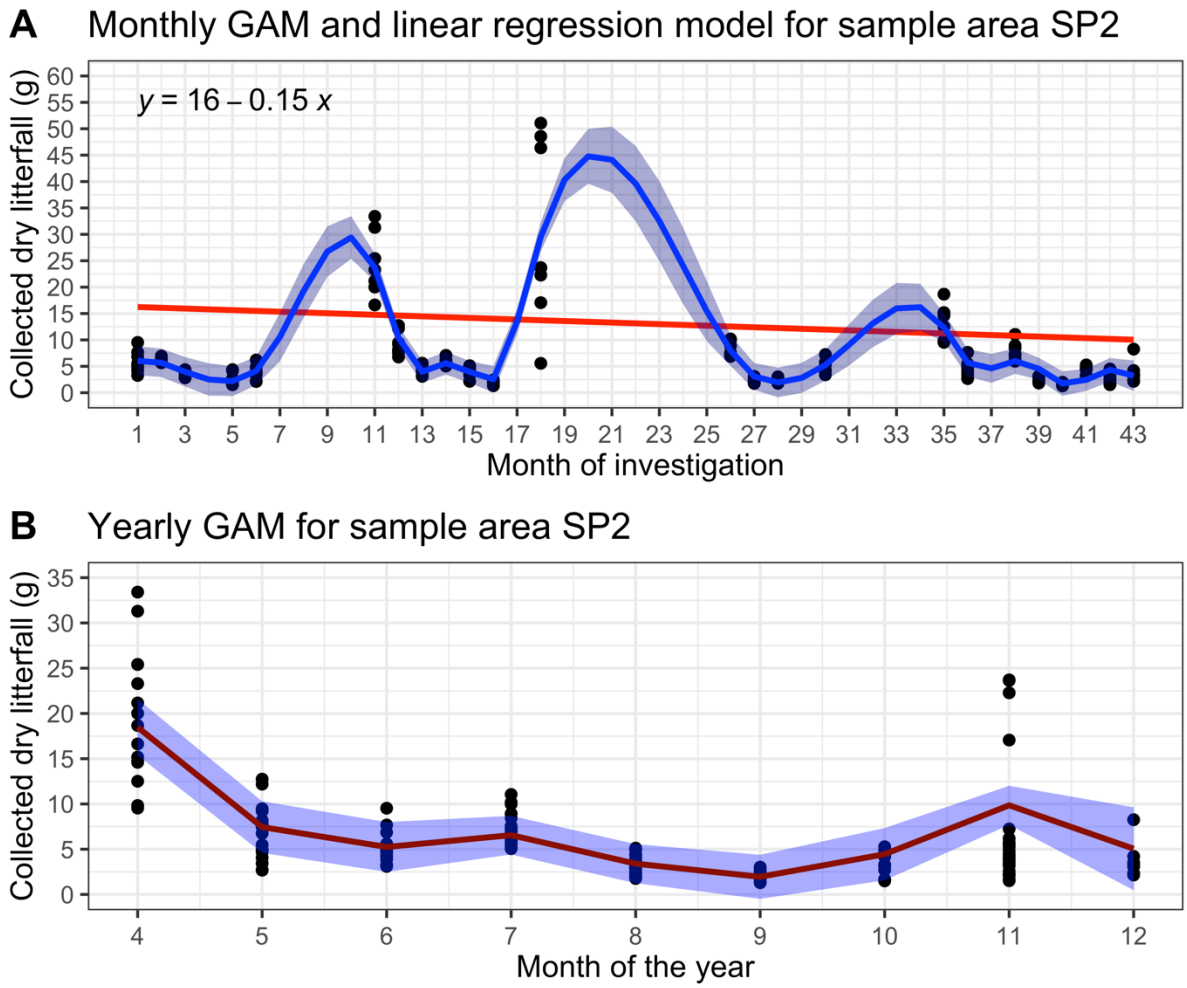


Figure 16. Monthly and yearly generalized additive models for sample area SP2.

Linear regression equation shows that amount of litterfall is slowly decreasing during four years of investigation. This leads to the conclusion that the litter amount in old forest stand is slowly decreasing. However, this equation could result from last autumn, when litterfall mass was smaller compared with the other three autumns.

The deviance explained parameter is 76,4, which is slightly higher than in other coniferous forests. In general, monthly GAM describes the litterfall dynamic on a decent level, but it is slightly overfitted.

The yearly generalized additive model demonstrates that litterfall accumulates quicker than in other coniferous sample areas during January, February and March. This observation negatively influences on prediction strength of different models. It is possible to assume that from January till March, the litterfall level will be on the same level as in May, but the absence of this data makes the models overfitted. Although, yearly generalized additive

model deviance explained parameter is only 34,7, which is not a significant result, but enough to show the basic relationships between variables.

The sample area SP3 tree composition is absolutely different from others. Sample area SP3 is a broadleaved forest. In the Figure 17 is represented monthly and yearly generalized additive models of this area.

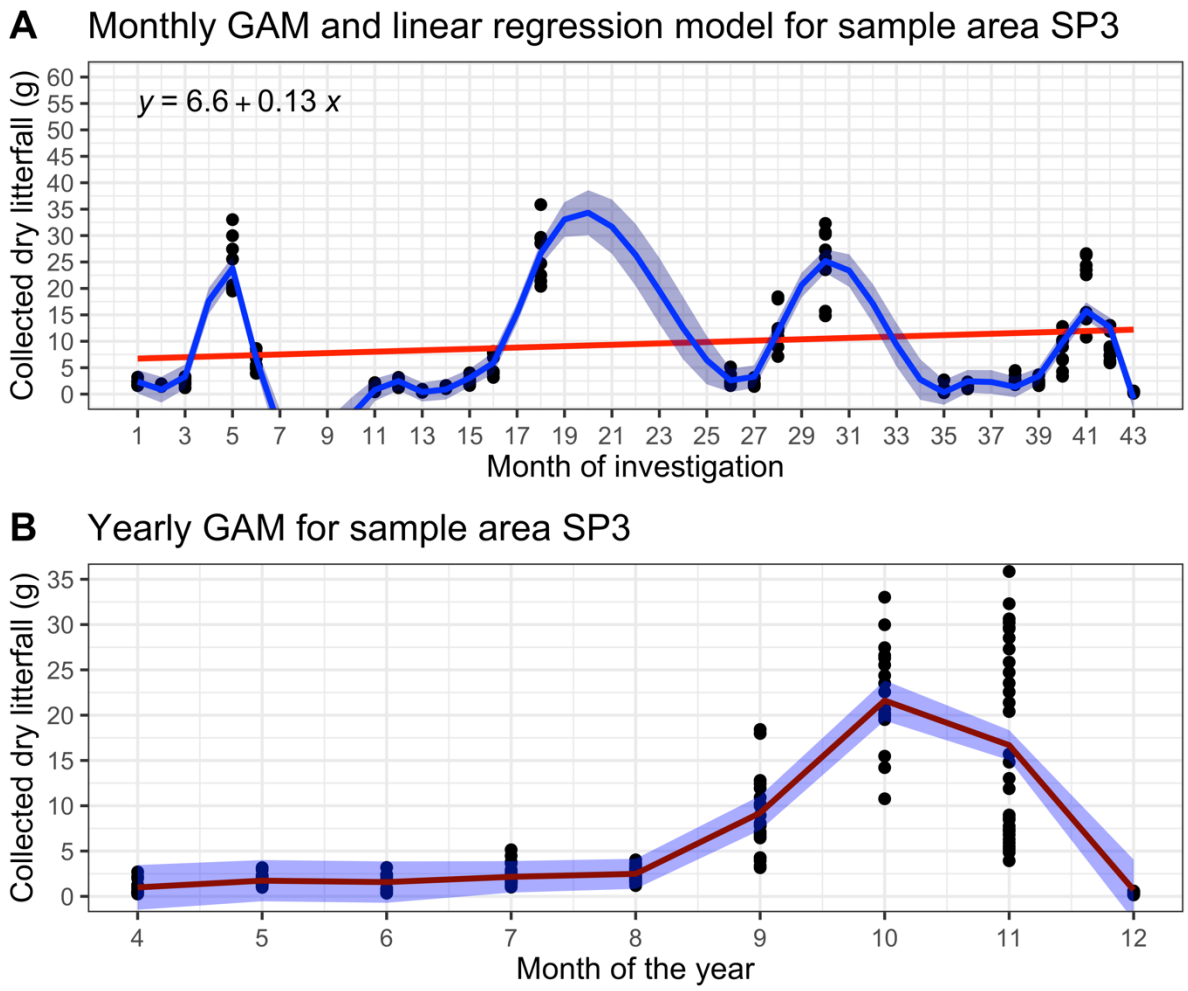


Figure 17. Monthly and yearly generalized additive models for sample area SP3.

The monthly generalized additive model perfectly demonstrates seasonal data pattern. Plot A excellently demonstrates four spikes and four crashes near or below zero values. These observations strongly demonstrate the natural litterfall process in autumn and the extremely low possibility of finding litterfall in litter trap in winter. Linear regression equation demonstrates a high growth of litterfall during all four years, expected from the young forest.

The deviance explained parameter is the highest from all four models – 88,1%. As a result, the monthly generalized additive model represents the litterfall dynamic very well.

The yearly generalized model is near to the real natural litterfall process. It perfectly shows the trees growth period from April till August and the litterfall period in October and November. Moreover, this model has the highest deviance explained parameter – 69,9%, which means that model fits the dataset very well. Overall, yearly GAM is showing basic relationships between variables with decent accuracy.

For solving the sampling problem, which is present in all previous models, a bootstrap method was used. In the Figure 18 is represented two models. The first model is the daily generalized additive model. The daily generalized additive model is the model, where the independent variable is the calendar day of the year, and the dependent variable is collected litterfall. In the Figure 18 it is represented by the solid blue line. The model's dark blue shaded area are the 95% confident intervals, and blue dots on the plot are actual litterfall observations during the four-year investigation. As mentioned in subchapter 1.4, outliers were not removed in this dataset. The second model is bootstrapped generalized additive model, represented by the solid red line. This is the same model as described in subchapter 2.6 in the Figure 11. Dark red shaded areas around the model are the confidence intervals, and the red dots are values generated by the bootstrap method. Y-axis is represented by different parameters. The y-axis represents the collected amount of litterfall in the daily general additive model. However, in the bootstrapped generalized additive model, the y-axis shows predicted values of litterfall. The X-axis represents the calendar day of the year for both models.

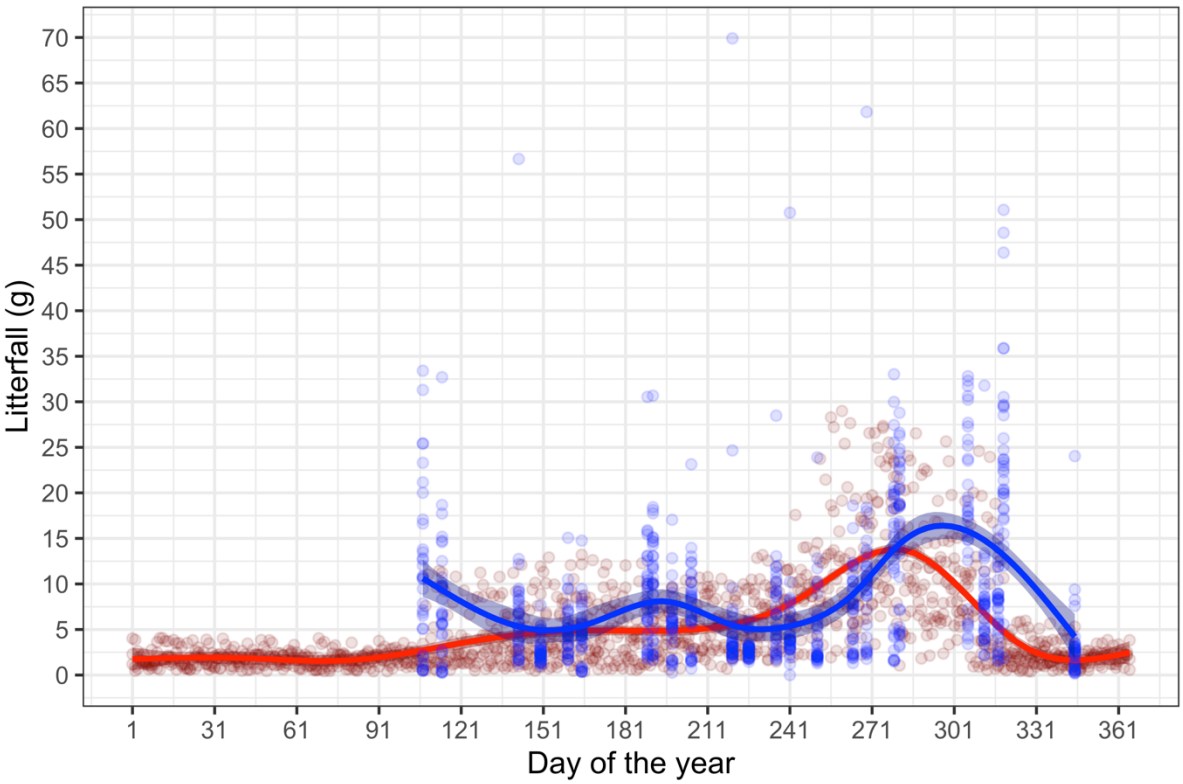


Figure 18. Compared daily generalized additive model and bootstrapped generalized additive models.

The most evident difference between these two models is the presence of daily values in January, February, March in the bootstrapped generalized additive model. In these three months, the flat segment of the model is observed. The reason is in the continuous uniform distribution. Values between late November or December and early April or May observations are uniformly distributed between the period in days. This period is relatively long, so values are relatively low. As a result, this also influences April values. Daily generalized model beginning from April and smoothly falls till June. This decrease does not reflect the actual process in the hemiboreal climate zone. Due to continuous uniform distribution, the inappropriate spike problem is solved.

The second important difference is a loss of a noticeable spike in late July. This spike could be explained by mathematical algorithms of distribution, which aligned this spike.

The third essential difference is the litterfall peak shift. The peak of litterfall is predicted in middle or late October in the daily generalized additive model. Due to uniform data distribution, this phenomenon is shifted back in late September or early October.

The last noticeable difference is the falling curve from November till December on the daily generalized additive model and the early-middle November curve on bootstrapped generalized additive model. Both of them are showing decreasing in litterfall quantity. However, daily GAM has the convex curve and bootstrapped GAM has the concave curve.

Although, it is incorrect to compare these two models using statistical parameters. Akaike criterion does not work in this case because the original dataset and generated dataset have different elements. Deviance explained parameter in the bootstrapped model is 49,8, which could be interpreted as 49,8% of dataset coverage. The result is significant. It does not overfit but clearly shows general relationships between variables.

4. CONCLUSION

To describe litterfall dynamics two pairs of linear regression and generalized additive models were generated and estimated. After comparing the linear regression models and the generalized additive models, it was concluded that choosing the rightly scaled independent variable is a highly important step. In this master's thesis the independent variable, which is the time axis, was represented in two ways. The first way demonstrated the litterfall process for four years, the second way demonstrated the litterfall as a cycling process. In the first case, the linear regression model was unsuitable to describe or predict the litterfall process. In the second case, the linear regression model described only the general trend during one calendar year based on the representative sample. As a result, both linear models were not good in describing or predicting detailed information.

In contrast, both generalized additive models deal with description or prediction tasks significantly better. Certainly, non-regular observations negatively influence the accuracy of models, but overall, their describing and predicting possibilities are greater.

Based on these results, it is possible to verify the first hypothesis of this master's thesis. The general additive model is a better choice in describing the litterfall process than the linear regression model.

Based on the first result, each sample area dataset was analysed with two types of generalized additive models. As previously, independent variable, which is the time axis, was represented in two ways.

Sample area A is represented by mixture of old and young coniferous forest stands. Both generalized additive model's prediction strength is harmed because of absent observation, but in general, gives the understanding of the basic relationships between variables.

Sample area SP1 is also represented by young coniferous forest with a similar species composition like sample area A. However, both general additive models are different from sample area A models.

Sample area SP2, what is represented by old coniferous forest stand. Here sample absence in wintertime plays the meaningful role. The Norway spruces (*Picea abies*) and other trees, what are growing in this area, accumulated a lot of litterfall during winter months. This leads to significant amount of litter in April, when the first collections of the investigation years were done. Both generalized additive models took April amount of litter into account, so both models are showing the quickly decreasing general of litterfall. Despite this result, is possible to assume, that in real life amount of litterfall.

Sample area SP3 is represented by very young broadleaved trees, where both generalized additive models very well reflecting the natural processes. Monthly GAM shows great seasonality and transition from one period to another. Yearly GAM shows expected litterfall process from broadleaved forest.

After summing everything up, it is possible to conclude, that second hypothesis of this master's work is verified. All four forest stands with different trees composition and ages have different litterfall generalized additive models.

For solving the problem of the absence of observations and increasing the model prediction strength, the bootstrap method was used. Compared with the previously generated models, bootstrapped litterfall GAM predicted the absent observations, removed the majority of the previous models inaccuracies and shifted some litterfall events back. This made bootstrapped generalized additive model very reliable in describing the litterfall process. At bootstrapped model is perfectly noticeable trees growing stage and litterfall peak in autumn time. Compared with previously generated models, bootstrapped model is more accurate and predicting observations more reliably. Based on these results, it is possible to conclude, that the third hypothesis of this master's thesis is verified. Non-regular and missing observations have a negative impact on model's prediction strength. Bootstrap is a good method, which allows to reduce the influence of absent observations and makes model more accurate and reliable.

REFERENCES

Acharya, A. S., Prakash, A., Saxena, P., & Nigam, A. (2013). Sampling: why and how of it? *Indian Journal of Medical Specialities*, *4*(2), 330–333. https://doi.org/10.7713/ijms.2013.0032

Aerts, R. (2006). The freezer defrosting: Global warming and litter decomposition rates in cold biomes. In *Journal of Ecology* (Vol. 94, Issue 4, pp. 713–724). https://doi.org/10.1111/j.1365-2745.2006.01142.x

Akaike, Hirotogu. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In E. Parzen, K. Tanabe, & E. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Akaike, Hirotugu. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332. https://doi.org/10.1007/BF02294359

Amemiya, T. (1983). Chapter 6 Non-linear regression models. In *Handbook of Econometrics* (Vol. 1, pp. 333–389). https://doi.org/10.1016/S1573-4412(83)01010-7

Barbur, V. A., Montgomery, D. C., & Peck, E. A. (1994). Introduction to Linear Regression Analysis. *The Statistician*, *43*(2), 339. https://doi.org/10.2307/2348362

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1989). The New S Language. *Biometrics*, *45*(2), 699. https://doi.org/10.2307/2531523

Berger, T. W., & Glatzel, G. (1994). Deposition of atmospheric constituents and its impact on nutrient budgets of oak forests (Quercus petraea and Quercus robur) in Lower Austria. *Forest Ecology and Management*, *70*(1–3), 183–193. https://doi.org/10.1016/0378-1127(94)90085-X

Bevans, R. (2020). *An introduction to the Akaike information criterion*. https://www.scribbr.com/statistics/akaike-information-criterion/

Bray, J. R., & Gorham, E. (1964). Litter Production in Forests of the World†. *Advances in Ecological Research*, *2*(C), 101–157. https://doi.org/10.1016/S0065-2504(08)60331-1

Brooks, D. G., Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1989). Akaike Information Criterion Statistics. *Technometrics*, *31*(2), 270. https://doi.org/10.2307/1268842

Cameron, A. C., & Windmeijer, F. A. G. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, *77*(2), 329–342. https://doi.org/10.1016/s0304-4076(96)01818-0

Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. In *Wiley Interdisciplinary Reviews: Computational Statistics* (Vol. 11, Issue 3, pp. 1–11). https://doi.org/10.1002/wics.1460

Celentano, D., Zahawi, R. A., Finegan, B., Ostertag, R., Cole, R. J., & Holl, K. D. (2011). Litterfall dynamics under different tropical forest restoration strategies in Costa Rica. *Biotropica*, *43*(3), 279–287. https://doi.org/10.1111/j.1744-7429.2010.00688.x

Chan, K. Y., Kwong, C. K., Dillon, T. S., & Tsim, Y. C. (2011). Reducing overfitting in manufacturing process modeling using a backward elimination based genetic programming. *Applied Soft Computing Journal*, *11*(2), 1648–1656. https://doi.org/10.1016/j.asoc.2010.04.022

Cheng, R. C. H., & Traylor, L. (1995). Non-Regular Maximum Likelihood Problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 3–24. https://doi.org/10.1111/j.2517-6161.1995.tb02013.x

Cousens, J. E., & Newbould, P. J. (1968). Methods of Estimating the Primary Production of Forests. *The Journal of Applied Ecology*, *5*(3), 745. https://doi.org/10.2307/2401650

Dhakal, C. (2019). Regression invented as statistics. *International Journal of Interdisciplinary Research and Innovations*, *6*, 1–5. https://www.researchgate.net/publication/326972988_REGRESSION_INVENTED_AS_STATISTICS

Edwards, W., Liddell, M., Franks, P., Nichols, C., & Laurance, S. (2017). Seasonal patterns in rainforest litterfall: Detecting endogenous and environmental influences from long-term sampling. *Austral Ecology*, *43*(2), 225–235. https://doi.org/10.1111/aec.12559

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Efron, Bradley, & Tibshirani, R. (1985). The Bootstrap Method for Assessing Statistical Accuracy. *Behaviormetrika*, *12*(17), 1–35. https://doi.org/10.2333/bhmk.12.17_1

Egger, A., & Carpi, A. (2008). Data Analysis and Interpretation. *Visionlearning*, *1*. https://www.visionlearning.com/en/library/Process-of-Science/49/Data-Analysis-and-Interpretation/154

Ewel, J. J. (1976). Litter Fall and Leaf Decomposition in a Tropical Forest Succession in Eastern Guatemala. *The Journal of Ecology*, *64*(1), 293–308. https://doi.org/10.2307/2258696

Ezhova, E., Ylivinkka, I., Kuusk, J., Komsaare, K., Vana, M., Krasnova, A., Noe, S., Arshinov, M., Belan, B., Park, S. Bin, Lavric, J. V., Heimann, M., Petäjä, T., Vesala, T., Mammarella, I., Kolari, P., Bäck, J., Rannik, U., Kerminen, V. M., & Kulmala, M. (2018). Direct effect of aerosols on solar radiation and gross primary production in boreal and hemiboreal forests. *Atmospheric Chemistry and Physics*, *18*(24). https://doi.org/10.5194/acp-18-17863-2018

Frost, J. (2017a). *Five Reasons Why Your R-squared can be Too High*. Statistics by Jim. https://statisticsbyjim.com/regression/r-squared-too-high/

Frost, J. (2017b). *How To Interpret R-squared in Regression Analysis*. Statistics by Jim. https://statisticsbyjim.com/regression/interpret-r-squared-regression/

Frost, J. (2018). *5 Ways to Find Outliers in Your Data*. Statistics by Jim. https://statisticsbyjim.com/basics/outliers/

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association*, *82*(398), 371–386. https://doi.org/10.1080/01621459.1987.10478440

Hastie, T., & Tibshirani, R. (1990). Generalized additive models. In *Generalized Additive Models* (1s edition). https://doi.org/10.1201/9780203753781

Hayes, A. (2020). *R-Squared Definition*. Investopedia. https://www.investopedia.com/terms/r/r-squared.asp

Henry, G. (2009). Practical Sampling. In L. Bickman & D. Rog (Eds.), *The SAGE Handbook of Applied Social Research Methods* (Second edi). https://doi.org/10.4135/9781483348858.n3

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, *76*(2), 297–307. https://doi.org/10.2307/2336663

Kallis, A., Rosin, K., Pärnpuu, P., Loodla, K., & Šišova, V. (2019). *100 aastat Eesti ilma (teenistust)*. https://www.ilmateenistus.ee/wp-content/uploads/2019/05/100_aastat_Eesti_ilma_teenistust.pdf

Kasturi, S. N. (2019). *Underfitting and Overfitting in machine learning and how to deal with it !!!* Towards Data Science.

Khuri, A. I. (2013). Introduction to Linear Regression Analysis, Fifth Edition by Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. *International Statistical Review*, *81*(2), 318–319. https://doi.org/10.1111/insr.12020_10

Klemmedson, J. O. (1987). Influence of Oak in Pine Forests of Central Arizona on Selected Nutrients of Forest Floor and Soil. *Soil Science Society of America Journal*. https://doi.org/10.2136/sssaj1987.03615995005100060039x

Krasnova, A., Kukumägi, M., Mander, Ü., Torga, R., Krasnov, D., Noe, S. M., Ostonen, I., Püttsepp, Ü., Killian, H., Uri, V., Lõhmus, K., Sõber, J., & Soosaar, K. (2019). Carbon exchange in a hemiboreal mixed forest in relation to tree species composition. *Agricultural and Forest Meteorology*, *275*, 11–23. https://doi.org/10.1016/j.agrformet.2019.05.007

Krishna, M. P., & Mohan, M. (2017). Litter decomposition in forest ecosystems: a review. In *Energy, Ecology and Environment 2* (pp. 236–249). https://doi.org/10.1007/s40974-017-0064-9

Kulmala, M., Ezhova, E., Kalliokoski, T., Noe, S., Timo, V., Annalea, L., Jari, L., Risto, M., Jaana, B., Petäjä, T., & Kerminen, V. M. (2020). CarbonSink+ — Accounting for multiple climate feedbacks from forests. *Boreal Enviromental Research*, 145–159. http://www.borenv.net/BER/archive/pdfs/ber25/ber25-145-159.pdf

Lebrija-Trejos, E., Pérez-García, E. A., Meave, J. A., Poorter, L., & Bongers, F. (2011). Environmental changes during secondary succession in a tropical dry forest in Mexico. *Journal of Tropical Ecology*, *27*(5), 477–489. https://doi.org/10.1017/S0266467411000253

Lian, Y., & Zhang, Q. (1998). Conversion of a natural broad-leafed evergreen forest into pure and mixed plantation forests in a subtropical area: Effects on nutrient cycling. *Canadian Journal of Forest Research*, *28*(10), 1518–1529. https://doi.org/10.1139/x98-173

Lillis, D. (2017). *Generalized Linear Models in R, Part 2: Understanding Model Fit in Logistic Regression Output*. The Analysis Factor.

Liu, C. J., Westman, C. J., & Ilvesniemi, H. (2001). Matter and nutrient dynamics of pine (Pinus tabulaeformis) and Oak (Quercus variabilis) litter in North China. *Silva Fennica*, *35*(1), 4–12. https://doi.org/10.14214/sf.599

Liu, C., Westman, C. J., Berg, B., Kutsch, W., Wang, G. Z., Man, R., & Ilvesniemi, H. (2004). Variation in litterfall-climate relationships between coniferous and broadleaf forests in Eurasia. *Global Ecology and Biogeography*, *13*(2), 105–114. https://doi.org/10.1111/j.1466-882X.2004.00072.x

Lodge, D. J., Scatena, F. N., Asbury, C. E., & Sanchez, M. J. (1991). Fine Litterfall and Related Nutrient Inputs Resulting From Hurricane Hugo in Subtropical Wet and Lower Montane Rain Forests of Puerto Rico. *Biotropica*, *23*(4), 336. https://doi.org/10.2307/2388249

Lonsdale, W. M. (1988). Predicting the amount of litterfall in forests of the world. *Annals of Botany*, *61*(3), 319–324. https://doi.org/10.1093/oxfordjournals.aob.a087560

Lopes, M., Araújo, V., & Vasconcellos, A. (2015). The effects of rainfall and vegetation on litterfall production in the semiarid region of northeastern Brazil. *Brazilian Journal of Biology*, *75*(3). https://doi.org/10.1590/1519-6984.21613

Lowman, M. D. (1988). Litterfall and Leaf Decay in Three Australian Rainforest Formations. *The Journal of Ecology*, *76*(2), 451–465. https://doi.org/10.2307/2260605

Martínez-Yrízar, A., & Sarukhán, J. (1990). Litterfall patterns in a tropical deciduous forest in Mexico over a five-year period. *Journal of Tropical Ecology*, *6*(4), 433–444. https://doi.org/10.1017/S0266467400004831

Microsoft Corporation. (n.d.). *DATEVALUE function*. https://support.microsoft.com/en-us/office/datevalue-function-df8b07d4-7761-4a93-bc33-b7471bbff252

Millar, C. S. (1974). Decomposition of Coniferous Leaf Litter. *Biology of Plant Litter Decomposition*, *1*, 105–128. https://doi.org/10.1016/b978-0-12-215001-2.50010-6

Minitab Inc. (2016). *Five Reasons Why Your R-squared Can Be Too High*. Minitab Blog. https://blog.minitab.com/en/adventures-in-statistics-2/five-reasons-why-your-r-squared-can-be-too-high

Noe, S. M., Niinemets, Ü., Krasnova, A., Krasnov, D., Motallebi, A., Kängsepp, V., Jõgiste, K., Hõrrak, U., Komsaare, K., Mirme, S., Vana, M., Tammet, H., Bäck, J., Vesala, T., Kulmala, M., Petäjä, T., & Kangur, A. (2015). SMEAR Estonia: Perspectives of a large-scale forest ecosystem– Atmosphere research infrastructure. *Forestry Studies*, *63*, 56–84. https://doi.org/10.1515/fsmu-2015-0009

Norden, N., Angarita, H. A., Bongers, F., Martínez-Ramos, M., Cerda, I. G. D. La, Van Breugel, M., Lebrija-Trejos, E., Meave, J. A., Vandermeer, J., Williamson, G. B., Finegan, B., Mesquita, R., & Chazdon, R. L. (2015). Successional dynamics in Neotropical forests are as uncertain as they are predictable. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(26), 8013–8018. https://doi.org/10.1073/pnas.1500403112

Novozhilov, Y. K., Rollins, A. W., & Schnittler, M. (2017). Ecology and Distribution of Myxomycetes. In *Myxomycetes: Biology, Systematics, Biogeography and Ecology* (pp. 253–297). https://doi.org/10.1016/B978-0-12-805089-7.00008-1

Piotrowski, A. P., & Napiorkowski, J. J. (2013). A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling. *Journal of Hydrology*, *476*, 97–111. https://doi.org/10.1016/j.jhydrol.2012.10.019

Poole, M. A., & O'Farrell, P. N. (1971). The Assumptions of the Linear Regression Model. *Transactions of the Institute of British Geographers*, *52*, 145. https://doi.org/10.2307/621706

Raich, J. W., & Schlesinger, W. . (1992). The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate. *Tellus B*, *44*(2), 81–99. https://doi.org/10.1034/j.1600-0889.1992.t01-1-00001.x

Scherer-Lorenzen, M., Bonilla, J. L., & Potvin, C. (2007). Tree species richness affects litter production and decomposition rates in a tropical biodiversity experiment. *Oikos*, *116*(12), 2108–2124. https://doi.org/10.1111/j.2007.0030-1299.16065.x

Shen, G., Chen, D., Wu, Y., Liu, L., & Liu, C. (2019). Spatial patterns and estimates of global forest litterfall. *Ecosphere*, *10*(2). https://doi.org/10.1002/ecs2.2587

Soetewey, A. (2020). *Outliers detection in R - Stats and R*. Stats and R. https://statsandr.com/blog/outliers-detection-in-r/

Song, P. X.-K. (2007). Correlated Data Analysis: Modeling, Analytics, and Applications. In *Correlated Data Analysis: Modeling, Analytics, and Applications*. https://doi.org/10.1007/978-0-387-71393-9

Starr, M., Saarsalmi, A., Hokkanen, T., Merilä, P., & Helmisaari, H. S. (2005). Models of litterfall production for Scots pine (Pinus sylvestris L.) in Finland using stand, site and climate factors. *Forest Ecology and Management*, *205*(1–3), 215–225. https://doi.org/10.1016/j.foreco.2004.10.047

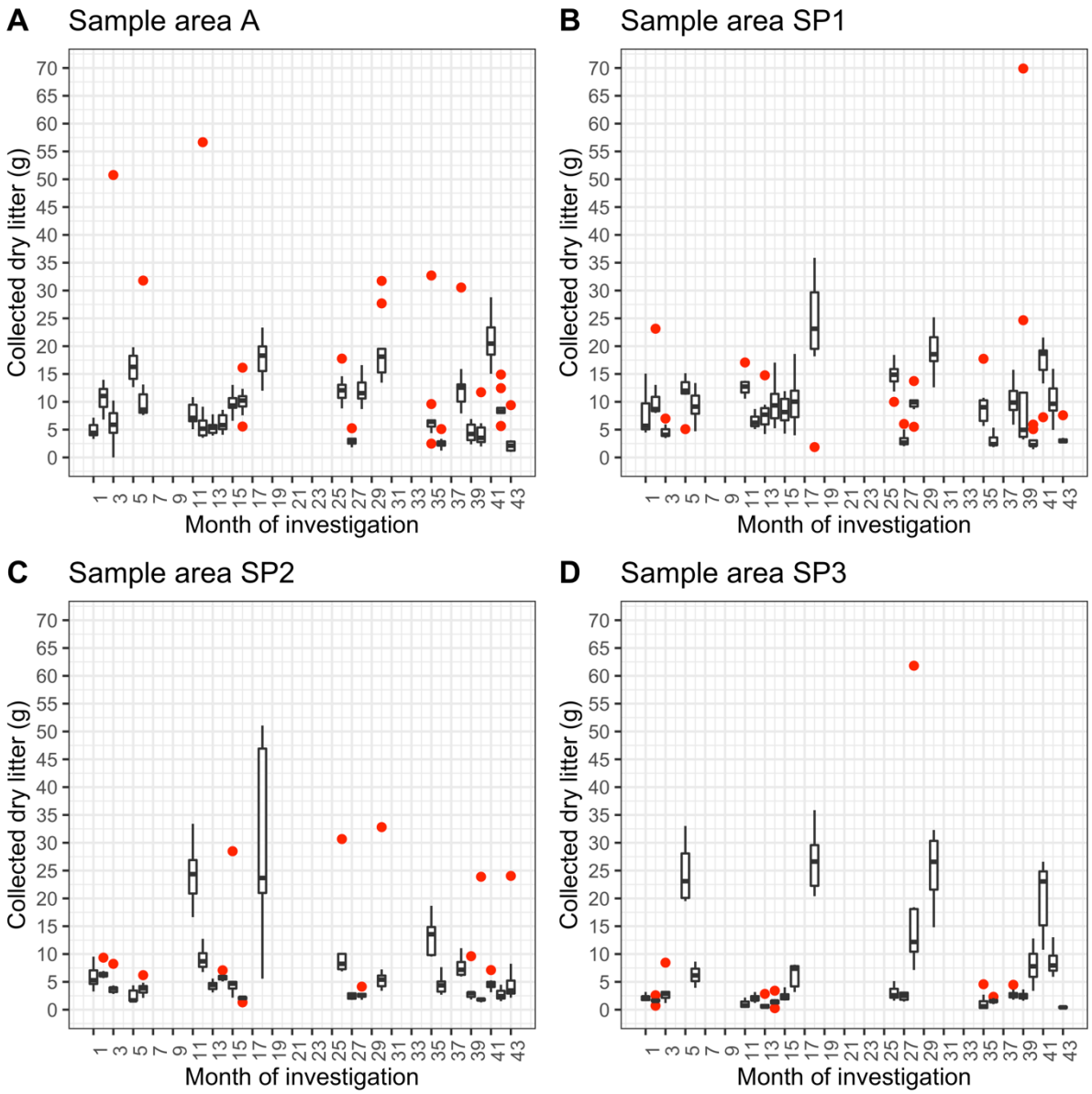
Tiit, E.-M. (2016). About the history of statistics. In *Quarterly Bulletin of Statistics 2/11*. https://www.stat.ee/dokumendid/55325

Van Der Aalst, W. M. P., Rubin, V., Verbeek, H. M. W., Van Dongen, B. F., Kindler, E., & Günther, C. W. (2010). Process mining: A two-step approach to balance between underfitting and overfitting. *Software and Systems Modeling*, *9*(1), 87–111. https://doi.org/10.1007/s10270-008-0106-z

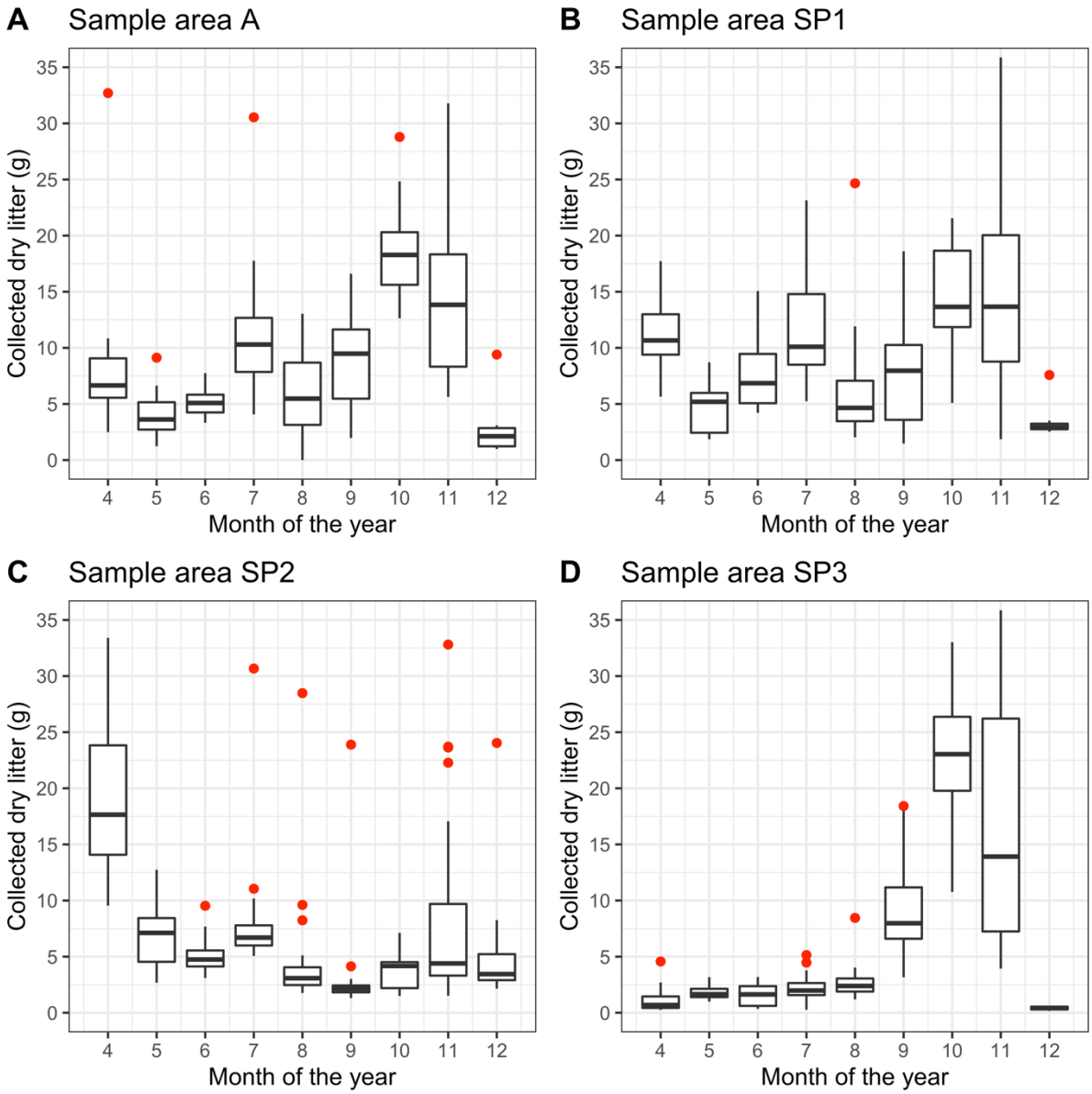
Varian, H. (2005). Bootstrap Tutorial. *Mathematica Journal*, *9*(4), 768–775.
Vellido, A., Martín-Guerrero, J. D., & Lisboa, P. J. G. (2012). Making machine learning models interpretable. *ESANN 2012 Proceedings, 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 163–172.
Vitousek, P. . (1982). Nutrient cycling and nutrient use efficiency. *American Naturalist*, *119*(4), 553–572. https://doi.org/10.1086/283931
Vitousek, P. M. (1984). Litterfall, nutrient cycling, and nutrient limitation in tropical forests. *Ecology*, *65*(1), 285–298. https://doi.org/10.2307/1939481
Wood, S. (2007). *Generalized Additive Models* (Second edi). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279
Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, *62*(4), 1025–1036. https://doi.org/10.1111/j.1541-0420.2006.00574.x
Xin, Y., & Su, X. G. (2010). Linear Regression Analysis - Theory and Computing. In *Linear Regression Analysis - Theory and Computing*. https://doi.org/10.1142/9789812834119
Yan, X. (2008). Linear Models: An Integrated Approach. *Journal of the American Statistical Association*, *103*(482), 884–885. https://doi.org/10.1198/jasa.2008.s234

APPENDIX

Appendix 1. Monthly collected litterfall box plots with detected outliers for each sample area



Appendix 2. Yearly collected litterfall box plots with detected outliers for each sample area



Lihtlitsents lõputöö salvestamiseks ja üldsusele kättesaadavaks tegemiseks ning juhendaja(te) kinnitus lõputöö kaitsmisele lubamise kohta

Mina, Svyatoslav Rogozin,
(sünnipäev 12/15/1996) 39612153716

1. annan Eesti Maaülikoolile tasuta loa (lihtlitsentsi) enda loodud lõputöö
Varise tekke ja dünaamika hindamine,
mille juhendaja on Steffen Manfred Noe,

   1.1. salvestamiseks säilitamise eesmärgil,
   1.2. digiarhiivi DSpace lisamiseks ja
   1.3. veebikeskkonnas üldsusele kättesaadavaks tegemiseks

kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile;

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Lõputöö autor allkirjastatud digitaalselt
allkiri

Tartu, 25.05.2021

Juhendaja(te) kinnitus lõputöö kaitsmisele lubamise kohta

Luban lõputöö kaitsmisele.

allkirjastatud digitaalselt	25.05.2021
_____	_____
(*juhendaja nimi ja allkiri*)	(*kuupäev*)
_____	_____
(*juhendaja nimi ja allkiri*)	(*kuupäev)*